An overview of clustering methods with guidelines for application in mental health research

Article *in* Psychiatry Research · May 2023

DOI: 10.1016/j.psychres.2023.115265

Caroline Gao
University of Melbourne
186 PUBLICATIONS **1,361** CITATIONS
SEE PROFILE

Kate Filia
Orygen The National Centre of Excellence in Youth Mental health
89 PUBLICATIONS **1,323** CITATIONS
SEE PROFILE

Johanna Bayer
Radboud University
30 PUBLICATIONS **699** CITATIONS
SEE PROFILE

Christoph Bergmeir
Monash University (Australia)
144 PUBLICATIONS **6,258** CITATIONS
SEE PROFILE

Psychiatry Research 327 (2023) 115265

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Review article

An overview of clustering methods with guidelines for application in mental health research

Caroline X. Gao[a,b,c,*], Dominic Dwyer[a,b], Ye Zhu[d], Catherine L. Smith[c], Lan Du[e], Kate M. Filia[a,b], Johanna Bayer[a,b], Jana M. Menssink[a,b], Teresa Wang[e], Christoph Bergmeir[e,f], Stephen Wood[a,b,1], Sue M. Cotton[a,b,1]

[a] Centre for Youth Mental Health, The University of Melbourne, Parkville, VIC, Australia
[b] Orygen, Parkville, VIC, Australia
[c] Department of Epidemiology and Preventative Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, VIC, Australia
[d] School of Information Technology, Deakin University, Geelong, VIC, Australia
[e] Faculty of Information Technology, Monash University, Clayton, VIC, Australia
[f] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

ARTICLE INFO

Keywords:
Clustering
Cluster analysis
Machine learning
Unsupervised learning
Mental health research

ABSTRACT

Cluster analyzes have been widely used in mental health research to decompose inter-individual heterogeneity by identifying more homogeneous subgroups of individuals. However, despite advances in new algorithms and increasing popularity, there is little guidance on model choice, analytical framework and reporting requirements. In this paper, we aimed to address this gap by introducing the philosophy, design, advantages/disadvantages and implementation of major algorithms that are particularly relevant in mental health research. Extensions of basic models, such as kernel methods, deep learning, semi-supervised clustering, and clustering ensembles are subsequently introduced. How to choose algorithms to address common issues as well as methods for pre-clustering data processing, clustering evaluation and validation are then discussed. Importantly, we also provide general guidance on clustering workflow and reporting requirements. To facilitate the implementation of different algorithms, we provide information on R functions and libraries.

1. Introduction

In the presence of the substantial variety of different ways humans can suffer with illnesses (Feczko et al., 2019; Nunes et al., 2020), there have been attempts to group individuals together who demonstrate similar aetiologies, presentations, prognoses, and responses to treatments. As highlighted by Sokal (1974), written history of classification schemes that attempt to categorise the natural world date back to at least the ancient Greeks. Throughout this time, grouping was achieved subjectively by determining similarities between organisms or objects using human sense perception; for example, birds that looked the same were grouped together. Such subjective methods of clustering continued with systematic biological taxonomies (e.g., Linnean typologies) and the emergence of nosologies in psychiatry (e.g., Kraepelinan dementia praecox).

With increasing amounts of data, a step-change in biological taxonomy was in the advent of computers that allowed simultaneous consideration of more variables than a human could and objective techniques to assess similarity, e.g., hierarchical agglomerative techniques and the numerical taxonomy approach of Sneath and Sokal (1973); Sokal and Sneath (1963). Computational classification also has a renewed interest in psychiatry due to the growing repositories of big data coupled with widespread recognition of multi-modal heterogeneity of individuals with the same diagnosis. For example, large variations in clinical presentations are well-recognized (Hyman, 2010), comorbidity occurs over the lifetime (Caspi et al., 2020), functional impairment widely varies (Dwyer et al., 2022), illness courses are unpredictable (Carpenter and Kirkpatrick, 1988), and there are multiple biological differences between individuals (Abi-Dargham and Horga, 2016; Insel and Cuthbert, 2015; Kapur et al., 2012) Simultaneously, there is recognition of the wide multi-modal similarity between individuals with different diagnoses in symptoms and biological measures, such as risk

* Corresponding author at: Centre for Youth Mental Health, University of Melbourne, Orygen, Locked Bag 10, 35 Poplar Road, Parkville, VIC 3052, Australia.
E-mail address: carolineg@unimelb.edu.au (C.X. Gao).

[1] Joint senior author

https://doi.org/10.1016/j.psychres.2023.115265
Received 15 December 2022; Received in revised form 20 May 2023; Accepted 21 May 2023
Available online 27 May 2023

genes (Lam et al., 2019; Schork et al., 2019; Zheutlin et al., 2019), blood proteins (Pinto et al., 2017), brain pathophysiology (Goodkind et al., 2015; Romer et al., 2021; Sha et al., 2019), cognition (Abramovitch et al., 2021), and more recent digital biomarkers (Fraccaro et al., 2019; Low et al., 2020).

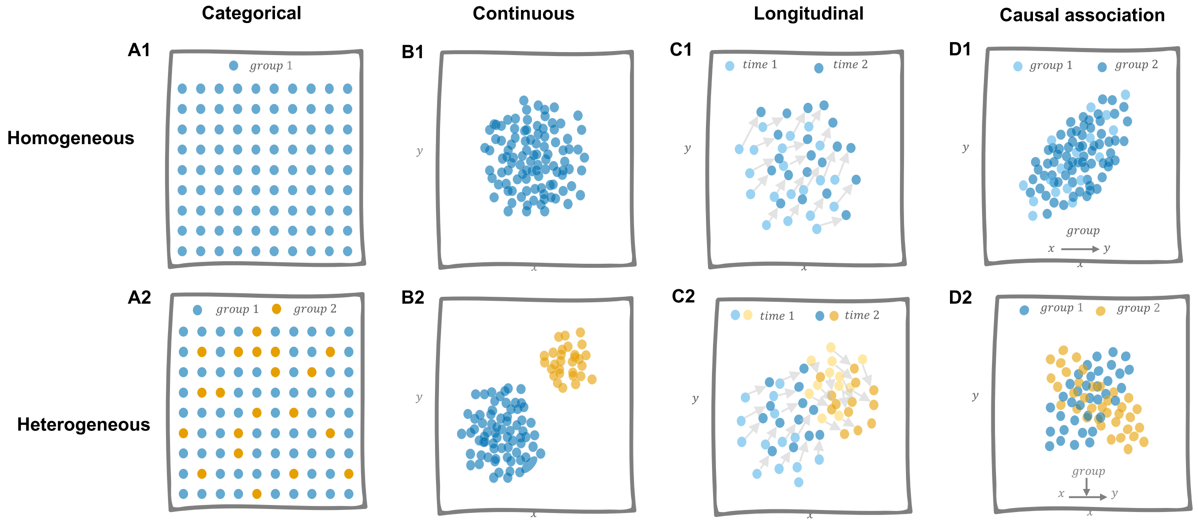
Such a growing array of differences and similarities has ignited debate within psychiatry that mirrors longstanding questions regarding the degree to which we either 'lump' or 'split' individuals into groups (McKusick, 1969). On one hand, there are entirely dimensional approaches that seek to identify major axes of illness variance shared by all individuals (Caspi et al., 2014; Caspi and Moffitt, 2018; Insel et al., 2010; Kotov et al., 2016, 2018, 2017, 2013) and on the other, there are approaches that attempt to discover subgroups of individuals who share distinctive attributes that separate them from others (Feczko et al., 2019). In this paper, we focus on the latter 'clustering' approach, but also note that the two extremes are not mutually exclusive and can greatly overlap and inform each other (Marquand et al., 2016).

Clustering, also referred to as cluster analysis (Feczko and Fair, 2020), is an unsupervised machine (statistical) learning technique that involves grouping data points together based on their similarities (outcomes or data labels are unknown, see Table S1 in Supplementary Material). The term "cluster analysis" was first used by Tryon (1939), and started to be implemented into computer algorithms in the 1960s, e.g., k-means clustering and hierarchical clustering (Forgy, 1965; Ward, 1963). Advances in machine learning in recent years have allowed clustering algorithms to be extended in functionality, scalability and complexity (Jain, 2010) to assist with understanding heterogeneity in mental health, see Fig. 1. A variety of clustering algorithms can now be found in most statistical packages such as R, Python, Matlab, Stata, SAS and IBM SPSS, and new algorithms continue to be developed and distributed rapidly, especially in R and Python.

Despite the increasing popularity of using clustering to identify homogeneous subgroups, there is little guidance on model choice, analytical frameworks and reporting requirements that links pioneering work in psychology (e.g.,Clatworthy et al. 2005; Milligan and Cooper 1987) with recent developments in machine learning (Jain, 2010; Russell and Norvig, 2021). Compared with commonly used statistical inference and prediction models, clustering tasks are often more challenging due to their explorative nature (Jain, 2010). Choices of algorithms and input parameters, statistical procedures, as well as randomness in parameter estimations, can all substantially influence the resulting groupings. Consequently, studies using these methods in mental health research often fail to demonstrate consistency (e.g., validation process), robustness (e.g., results not impacted by randomness in estimation) and reproducibility (e.g., details in reporting, open access code and data) (Clatworthy et al., 2005; Green et al., 2020; Ulbricht et al., 2018; Zhou et al., 2018). Therefore, there is a need to provide a comprehensive understanding of common clustering models and to establish frameworks for conducting cluster analysis.

There have been many reviews of existing clustering algorithms (Ezugwu et al., 2020; Feczko and Fair, 2020; Gan et al., 2020; Jain, 2010; Jain et al., 1999b; Rui and Wunsch, 2005; Xu and Tian, 2015). However, these reviews generally target statistics, machine learning and computer science audiences. Most of these reviews only provide a general summary of methods without detailed practical guidance for those less familiar with these approaches. Thus, the aim of this paper is to: (i) provide an overview (the philosophy, design and implementation) of major clustering methods that are particularly relevant in mental health research; (ii) introduce the extensions of basic models; (iii) discuss important issues commonly faced in clustering tasks; and (iv) provide general guidance on the clustering workflow.

Given the increasing popularity of the statistical computing software R in mental health research, we also provide information on R functions and libraries for implementing different algorithms. The paper is



Heterogeneity in mental health: refers to variations between individuals in clinical symptoms, severities, trajectories, causal pathways, environmental and physiological/biological factors associated with mental health conditions. In practice, it is frequently observed in:

(A) distinct groups such as different diagnoses, e.g., A1 – anxiety only (group1) vs A2 – anxiety (group1) and depression (group 2);
(B) different severity levels for clinical symptoms, e.g., B1 – moderate level of anxiety (x) and depression (y) vs B2 – one group with mild anxiety (x) and depression (y) and the other with severe anxiety (x) and depression (y);
(C) different treatment trajectories, e.g., C1 – improvement in both quality of life (x) and function (y) over time vs C2 – one group with improvement in quality of life (x) and function (y) and the other with improvement in quality of life (x) but a reduction in function (y);
(D) different causal pathways, e.g., D1 – Group 1 and 2 had similar treatment (x) and outcome (y) response vs D2 – group 1 with improvements in outcome (y) with treatment and group 2 with a reduction in outcome (y) with treatment.

Fig. 1. Heterogeneity in Mental Health.

targeted at readers in mental health research, but the analytical content is applicable to other research fields such as public health and social science.

2. Clustering algorithms

Thousands of clustering algorithms have been published with variations in their fundamental design, assumptions, target data structures, parameters of interest, and computational/optimisation processes. The taxonomy of clustering algorithms can be mapped out in different ways (Fig. S1 in Supplementary Material), for example, partitional vs hierarchical, or soft vs hard (Giordani et al., 2020; Han et al., 2011; Reddy and Vinzamuri, 2018). Considering the common problems encountered in mental health research, we broadly summarize the common algorithms into four groups, namely center-based partitioning clustering, hierarchical clustering, density-based clustering, and model-based clustering (Fig. 2).

2.1. *Similarity and dissimilarity (distance) measures*

Most clustering algorithms (except for model-based clustering) require measuring similarity or dissimilarity to group observations into alike subgroups. There are many types of similarity and dissimilarity measures. Distance, sometimes used interchangeably with dissimilarity, is a commonly used sub-type of dissimilarity measures. The most straightforward method for measuring distance for continuous variables is the Euclidean distance. It is simply the length of the path connecting two points in two-dimensional space. Although the Euclidean distance is easy to calculate and interpret, it may not be suitable or optimal for the data given. It cannot be used when the variable is not continuous or when variables have scale differences. It is also sensitive to outliers. A range of different types of measures have been developed to address different data types and issues (Alamuri et al., 2014; Cha, 2007). Detailed descriptions of common distance or dissimilarity measures are listed in Table 1 (continuous variables) and Table 2 (binary, nominal or mixed variables).

The appropriate measure should be chosen according to the requirement of the clustering algorithm, the type of data (continuous, ordinal, nominal, binary, count or mixed), and whether the data have outliers (e.g., Manhattan distance is less sensitive to outliers compared with Euclidean distance). In some cases, the scale of the variable may not be associated with underlying differences (e.g., whether a word is presented in a document is more important than the frequency of the word). In these instances, measures such as Cosine distance should be used, which represents distance by the angle between vectors, therefore is scale-invariant (see Table 1). When data are sparse and have a high proportion of common zeros that do not represent similarity (for binary data), distance measures such as Hamming distance (Manhattan distance for binary data) can be problematic. This issue is also known as the "double-zero problem" (Legendre and Legendre, 2012). For example, when many clinical symptoms are measured, an individual presenting with only anxiety symptoms (anxiety=1, other symptoms=0,0,0,0…), should be considered very differently from another person presenting with only psychotic symptoms (psychotic=1, other symptoms=0,0,0, 0…). In this case, normalised measures such as Jaccard distance should be used as common 0 s are not included in the calculation of distance (see Table 2 for details).

2.2. *Common clustering algorithms*

Although most algorithms use distance measures, they were commonly designed with different philosophical rationales.

Center-based partitioning clustering (also known as center-based, distance-based or partitioning clustering), refers to a family of models

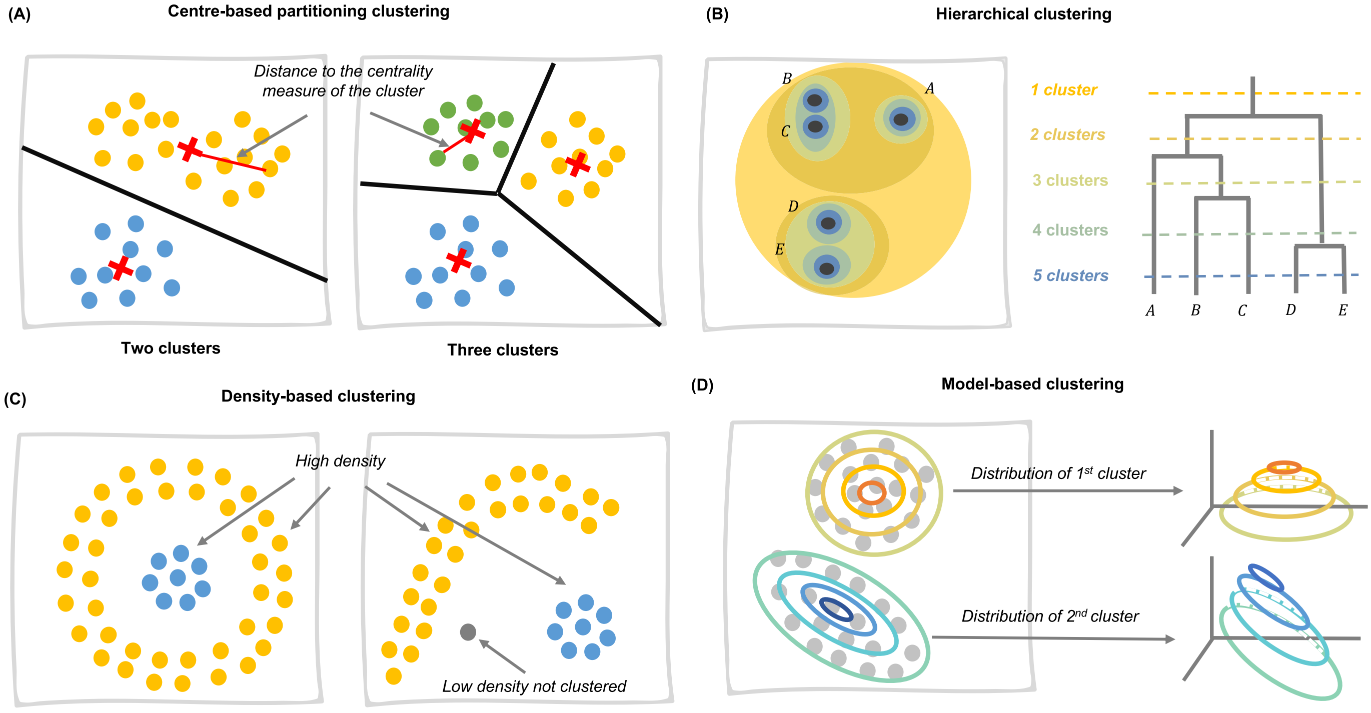
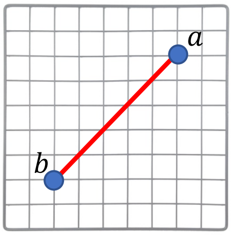
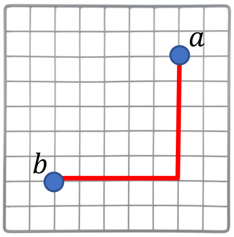


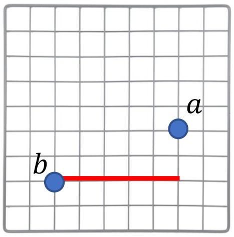
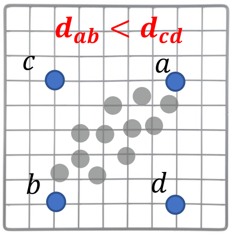
Fig. 2. Simple Illustration of Main Types of Clustering Models.
Note: (A) Center-based partitioning clustering aims at establishing the center of each cluster (with the number of clusters pre-specified) and determining group membership using the distance to the individual cluster center. (B) Hierarchical clustering groups data objects into a hierarchy or "tree" of clusters. (C) Density-based clustering groups data according to the density of data distribution, therefore, can identify clusters with arbitrary shapes and sizes. (D) Model-based clustering assumes the distribution of the data is underpinned by latent subgroups.

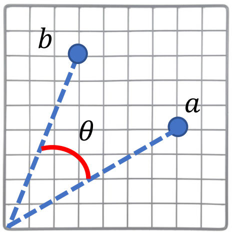
Table 1
Common Distance or Dissimilarity Measures for Continuous Variables.

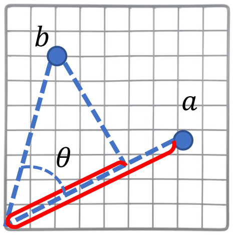
Distance or dissimilarity	Equation (distance between *a* and *b* on *n* dimensions)	Graphical representation	Description	R code

Euclidean distance[*]	$\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$		Distance between two points in *n* dimensional space. It is the most commonly used distance measure in clustering; however, it can be sensitive to outliers (Jain et al., 1999a). A few other distance measures were modified based on Euclidean such as weighted Euclidean distance, and average Euclidean distance.	stats::dist(x, method = "euclidean")
				philentropy::distance(x, method = "euclidean")

Manhattan distance[*]	$\sum_{i=1}^{n}\|a_i - b_i\|$		Calculated as the sum of the absolute differences in each dimension. It is faster to calculate and slightly more robust to outliers compared with Euclidean distance because there are no squared terms (Rui and Wunsch, 2005).	stats::dist(x, method = "manhattan")
				philentropy::distance(x, method = "manhattan")

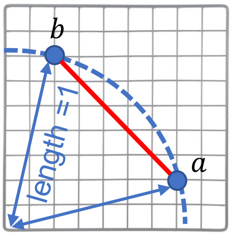
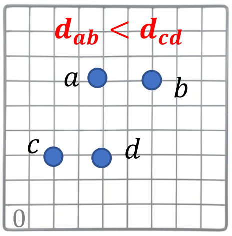
Chebyshev distance[*]	$\max_{i}\|a_i - b_i\|$		It measures the maximum distance on all given dimensions. It is very fast to calculate, however, might not be accurate because the information on other dimensions is suppressed.	philentropy::distance(x, method = "chebyshev")
Mahalanobis distance	$\sqrt{(\vec{a} - \vec{b})S^{-1}(\vec{a} - \vec{b})^T}$ *S* is the covariance matrix of all the dimensions in the data		Mahalanobis distance takes the data variation and correlation into account. Therefore, data points further away from their expected association were considered more dissimilar. Although this distance measure corrects for the data correlation structure, it is more computationally intensive and can have numerical issues when variables are highly correlated (De Maesschalck et al., 2000).	stats::mahalanobis(x, center = FALSE, cov = cov(x))

Cosine distance[^]	$\cos(\theta) = \frac{\sum_{i=1}^{n}a_i b_i}{\|\|a\|\|_2\ \|\|b\|\|_2}$
	$\|\|a\|\|_2 = \sqrt{\sum_{i=1}^{n}a_i^2}$
			Instead of geometrical distance, cosine measures the angle between vectors. It is useful when the actual value of data (such as frequency of words in documents) can be a biased representation of the underlying feature, which is common in text mining (Huang, 2008).	lsa::cosine(x)
				philentropy::distance(x, method = "cosine")

Dot product[^]	$\sum_{i=1}^{n}a_i b_i$ $= \|\|a\|\|_2\|\|b\|\|_2\cos(\theta)$		The dot product is similar to cosine similarity except that it considers the magnitude of vectors. The dot product can be very helpful when both the angle and the magnitude are important. For example, if we are interested in clustering frequencies developing clinical symptoms, both the frequency and symptom overlaps became important in defining how similar two patients are.	corrr::colpair_map(as.data.frame(t(x)), function(x, y) x%*% y,. diagonal=apply(x,1,function(x) x %*% x))

(*continued on next page*)

Table 1 (*continued*)

Distance or dissimilarity	Equation (distance between a and b on n dimensions)	Graphical representation	Description	R code
Chord distance[^]	$\sqrt{\sum_{i=1}^{n}\left(\frac{a_i}{\|\|a\|\|_2}-\frac{b_i}{\|\|b\|\|_2}\right)^2}$		Chord distance is the Euclidean distance with raw data normalised between 0 and 1 (relative to all variables collected for the data point, also known as chord transformation). The normalisation removes the impact of the data scale differences, and deal with a numerical problem known as “Double-zero problem” in ecology: when a value of 0 does not represent similarity (e.g., absence of a species), Euclidean distance cannot properly represent similarity (Legendre and Legendre, 2012).	stats::dist(vegan::decostand(x, method = "normalize"), method = "euclidean")
Canberra distance	$\sum_{i=1}^{n}\frac{\|\boldsymbol{a_i}-\boldsymbol{b_i}\|}{\|\boldsymbol{a_i}\|+\|\boldsymbol{b_i}\|}$		Manhattan distance is weighted by the inverse of the sum of absolute value, so the data is more sensitive to differences that are closer to 0.	philentropy::distance(x, method = "canberra")

[*] Euclidean, Manhattan and Chebyshev distance are special types of Minkowski distance, which has a general expression: $[\sum_{i=1}^{n}|a_i - b_i|^p]^{1/p}$

[^] $||a||_2$, also known as l_2-norm, is calculated as the square root of the sum of the squared, which represents the vector length. When using it as a normalising constant, it removes the impact of the vector length, e.g., two 2-dimensional vectors [0,1] and [10,10] became identical after dividing by their l_2-norm: $||[0,1]||_2 = [0,1\,/\sqrt{0^2+1^2}\,] = [0,1]$ and $||[0,10]||_2 = [0,10\,/\sqrt{0^2+10^2}\,] = [0,1]$

including K-means and related algorithms. The core concept of these models is to find mutually exclusive clusters with spherical shapes based on data points’ distance to cluster centers. K-means, discovered independently by different authors in the 1950s and 1960s (Ball and Hall, 1965; MacQueen, 1967), is perhaps the most popular clustering algorithm. As shown in Fig. 3, the algorithm iteratively estimates the centroids of clusters (based on Euclidean distances) until convergence. Although the algorithm has high computational speed and easy interpretation, K-means suffers from a few major limitations, including being sensitive to outliers, unable to identify non-spherical shapes, and only obtaining the local optimum solution (sensitive to the random starting point)(Jain, 2010). A range of algorithms was subsequently developed to address issues of K-means, for example, K-medoids clustering which uses medoids (the data point with the lowest average distance to all other points in the cluster) to reduce the model’s sensitivity to outliers; K-means++ to initialize centroids (Arthur and Vassilvitskii, 2006); fuzzy C-means, which assign probabilistic membership to account for overlapping clusters (Bezdek, 1981); K-modes for categorical input data (Huang, 1997b) and; K-prototypes for mixed numeric and categorical data (Huang, 1997a).

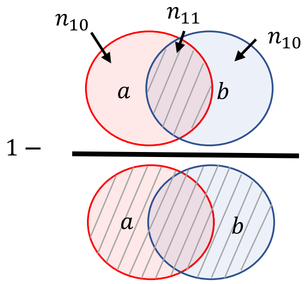
Hierarchical clustering takes a different approach to segmenting the data compared with center-based methods. The motivation for hierarchical clustering originated from the need for classification in biological taxonomy in the 1950s and 1960s (Johnson, 1967). It applies a hierarchical approach to establish a tree of clusters (as a dendrogram) either using a bottom-up (agglomerative) or a top-down (divisive, less popular due to high computational cost) based on distances between sub-clusters (Fig. 4). After completion of the algorithm, group membership of any given number of clusters can be determined by slicing the dendrogram. Hierarchical clustering can work with any similarity/dissimilarity matrix and has direct hierarchical interoperation of clusters using a dendrogram. As hierarchical clustering can be used to understand groupings and hierarchical structures of variables (using correlation coefficients as similarity measures), the model can fit well with the framework of accessing the hierarchical taxonomy of disorders and symptoms, for example, the Hierarchical Taxonomy of Psychopathology (HiTOP) (Kotov et al., 2017). However, it has a few drawbacks including being computationally expensive with large datasets, lacking the ability to predict cluster membership with new data, and lower level of stability (sensitive to minor data perturbation) (Milligan, 1980).

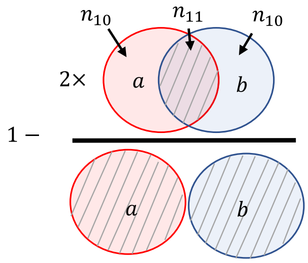
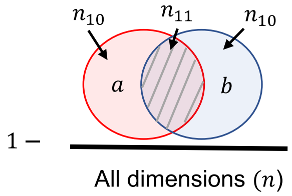
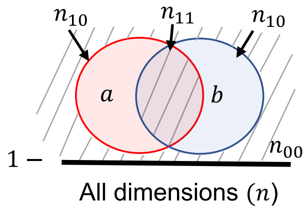
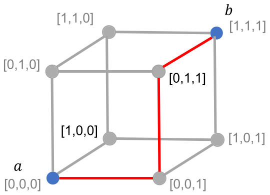
Density-based clustering is designed from a different school of thought, which aims at finding high-density areas in the feature space (vector space of all variables). The general process is to link (or grow) neighboring dense points to form clusters and leave points that are far away from dense regions as un-clustered. The most well know density-based model is DBSCAN proposed by Easter et al. (1996a), see illustration in Fig. 5. The benefits of DBSCAN include its robustness to outliers and noise in the data, and the ability to detect clusters with arbitrary shapes and sizes. However, it has two main limitations: it cannot work with clusters that have different densities and it is time-consuming to execute on large and high-dimensional data. These limitations motivated the development of extended models such as HDBSCAN (Campello et al., 2013). Another novel idea is Density Peak Clustering proposed by Rodriguez and Laio (2014), in which a user can decide the number of clusters by exploring locally densest data points as possible cluster centers. A comprehensive survey of density-based clustering algorithms is provided by Bhattacharjee and Mitra (2020). Density-based clustering can have significant benefits when researchers aimed at establishing clear boundaries between subclusters. For example, when aiming to identify subgroups of depression with distinct causal mechanisms, the researchers should avoid separating subgroups representing a continuum (e.g., low, medium, and high severity). In a re-evaluation of clustering for major depressive disorder by Drysdale et al. (2017), Dinga et al. (2019) found that the four cluster solution can also be replicated from data sampled from a single Gaussian distribution suggesting no clear boundaries between sub-biotypes.

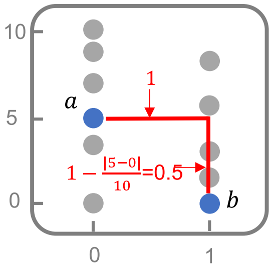
Another major approach in clustering analysis is model-based clustering, based on probability models. The most popular method in model-based algorithms is the finite mixture model. The finite mixture model was first applied in parameter estimation by Pearson (1894) and has been widely used in the scientific literature. In the 1960s and 1970s, finite mixture models began to be used for clustering problems (Day, 1969; Edwards and Cavalli-Sforza, 1965). The method has gained increasing popularity following the development of the

Table 2
Common Distance or Dissimilarity Measures for Binary, Nominal or Mixed Variables.

Distance or dissimilarity	Equation (distance between a and b on n dimensions)	Graphical representation	Description	R code

Jaccard distance	$1-\frac{a\cap b}{a\cup b}$
	For binary data:
	$1-\frac{n_{11}}{n_{11}+n_{01}+n_{10}}$
		 n_{10}
		n_{11}
		n_{10}
		a
		b
		1 −
		a
		b
			Calculated as the proportion of mismatches between two data points (range from 0 to 1), which indicate how diverse the two data points are. It is mostly used for binary data or unstructured normal data (text data). It gives more weight to common ones than common zeroes (Leisch, 2006), e.g., the distance between [1,0] and [1,1] are the same as between [1,0, 0,0,0] and [1,1,0,0,0] as the three common 0 dimensions are not included in the calculation.	stats::dist(x, method = "binary")
				prabclus::jaccard(t(x))
				philentropy::distance(x, method = "jaccard")
				proxy::dist(x, by_rows = TRUE, method = "Jaccard")

Dice distance	$1-\frac{2^*a\cap b}{a+b}$ For binary data: $1-\frac{2n_{11}}{2n_{11}+n_{01}+n_{10}}$	 n_{10} n_{11} n_{10} 2× a b 1 − a b	Modified version of Jaccard with more weights given to cases with agreements (common ones).	philentropy::distance(x, method = "dice")
Russell/Rao distance	$1-\frac{a\cap b}{n}$ For binary data: $1-\frac{n_{11}}{n}$ total dimensions $n=n_{11}+n_{01}+n_{10}+n_{00}$	 n_{10} n_{11} n_{10} a b 1 − All dimensions (n)	The proportion of disagreement between data points. Different from Jaccard and Dice distance, the common zeros are included in the calculation (n_{00} is included in the denominator). The inclusion of common zeros can be problematic when the matrix is sparse, the distance will become too large to identify differences with rare cases.	Mercator::binaryDistance(t(x), metric="russellRao")
Simple matching distance	$1-\frac{a\cap b+\bar{a}\cap\bar{b}}{n}$ For binary data: $1-\frac{n_{00}+n_{11}}{n}$	 n_{10} n_{11} n_{10} a b n_{00} 1 − All dimensions (n)	Simply calculated as the proportion of mismatched records (both common ones and common zeros) among all records. Similar to the Russell/Rao distance that cannot detect differences in sparse data.	nomclust::sm(x)
Hamming distance	$\sum_{i=1}^{n} a_i \neq b_i$	 b [1,1,0] [1,1,1] [0,1,0] [0,1,1] [1,0,0] [1,0,1] a [0,0,0] [0,0,1]	Also known as the Manhattan distance for binary data. It is calculated as the number of values that are different between two data points. It is easy to compute, but it does not consider whether the data consistency is related to common zeros or common ones. Therefore, it has similar issues as the Russell/Rao distance when dealing with a high dimensional sparse matrix (Norouzi et al., 2012). Hamming distance can also be used for nominal data, which counts for total mismatches regardless of the meaning or prevalence of each choice.	Mercator::binaryDistance(t(x), metric="hamming")

Gower distance	$1-\frac{1}{\sum_{i=1}^{n} w(a_i,b_i)}\times$
	$\sum_{i=1}^{n}\frac{1}{w(a_i,b_i)}s(a_i,b_i)$
	$s(a_i,b_i)$ is the similarity function and $w(a_i,b_i)$ is a weight *
		 10
		5
		0
		0
		1
		1
		a
		b
		$1-\frac{\lvert 5-0\rvert}{10}$=0.5
			Measures how different two data points are when mixed data (binary, categorical or continuous data) are presented. It is one of the most widely applied distance measures with mixed types of data. Weight can be applied (e.g., the weight of 0 for double zeros in binary data) to minimize the impact of the high dimensional sparse matrix (Gower, 1971). It can be sensitive to outliers.	cluster::daisy(x, metric="gower")
				StatMath::gower.dist(x, metric="gower")

[*] When i dimension is binary or categorical data when $a_i = b_i$ $s(a_i, b_i) = 1$, otherwise $s(a_i, b_i) = 0$. When i dimension is continuous, range normalised range-normalized Manhattan distance is used, $(a_i, b_i) = 1-\frac{|a_i-b_i|}{R_i}$, where R_i is the range of the i dimension.

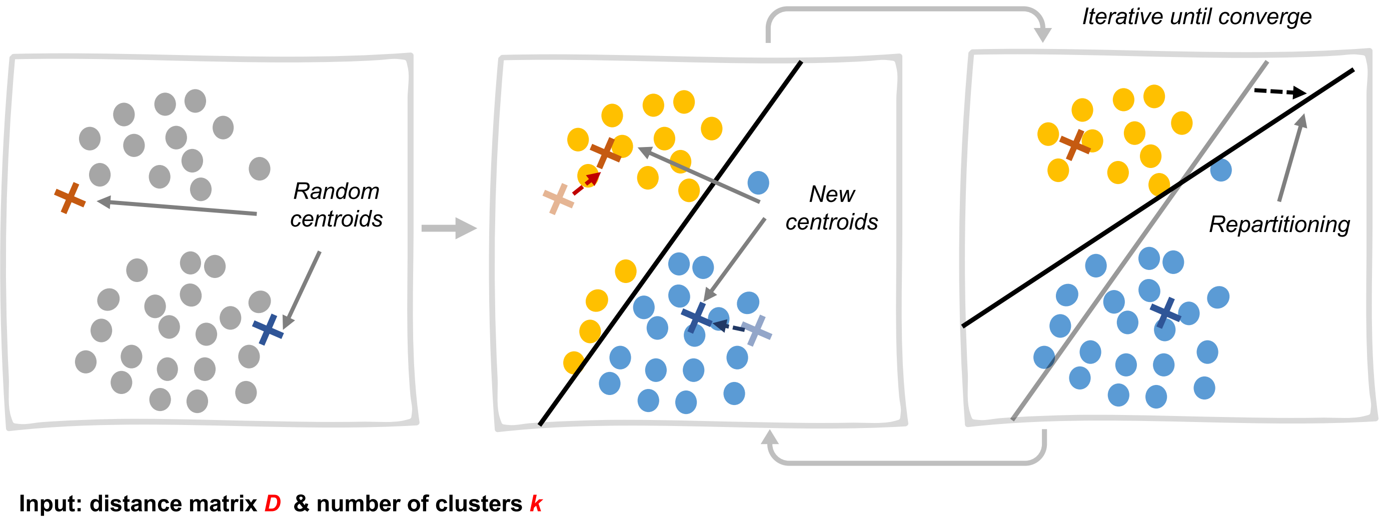


Fig. 3. Illustration of K-means Clustering.
Note: The estimation routine of K-means involves: (i) randomly initialise centroids of a pre-specified number of clusters and partition data points into groups according to their distance to the centroids; (ii) re-estimate the centroids (calculated as the mean of all the data points of the cluster) using points for each cluster; (iii) re-partition data into clusters; and (iv) iteratively repeat (ii) and (iii) until no more changes were observed in the location of centroids (convergence).

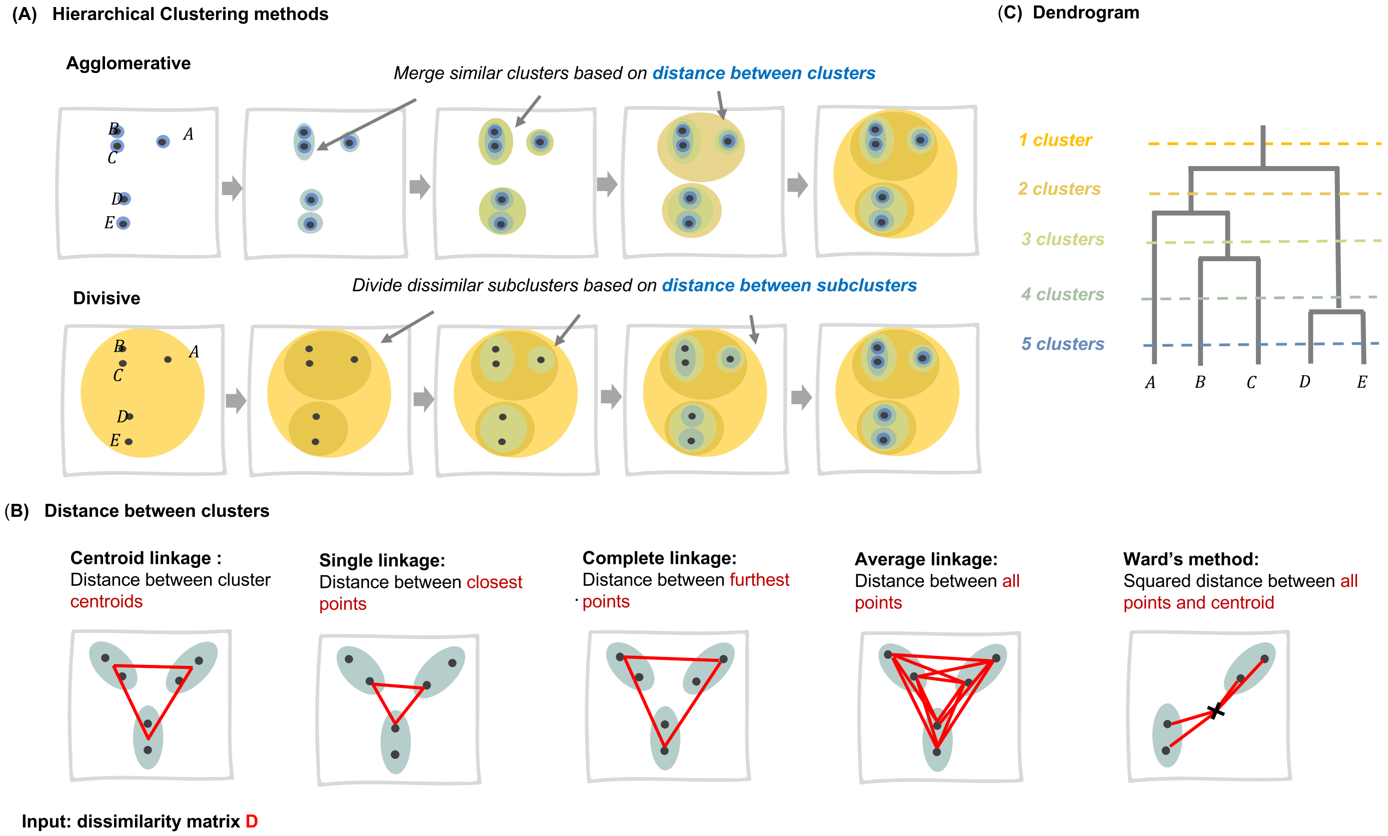

(A) Hierarchical Clustering methods
Agglomerative
Merge similar clusters based on distance between clusters
A
B
C
D
E
Divisive
Divide dissimilar subclusters based on distance between subclusters
(C) Dendrogram
1 cluster
2 clusters
3 clusters
4 clusters
5 clusters
A B C D E
(B) Distance between clusters
Centroid linkage : Distance between cluster centroids
Single linkage: Distance between closest points
Complete linkage: Distance between furthest points
Average linkage: Distance between all points
Ward's method: Squared distance between all points and centroid
Input: dissimilarity matrix D

Fig. 4. Hierarchical Clustering.
Note: (A) The agglomerative method is a bottom-up approach, which starts with treating individual data points as separate clusters and then iteratively merges “similar” clusters into larger clusters until all the data are in one cluster. The divisive hierarchical clustering, on the contrary, is a top-down approach, which starts with one cluster and sub-divides the cluster into smaller clusters. (B) Distance between subclusters can be measured using centroid linkage, single linkage, complete linkage, average linkage and Ward’s method (C) Dendrogram generated after clustering allows users to slice the hierarchical structure into any number of clusters.

Expectation-Maximization (EM) algorithm, an efficient algorithm used to find maximum likelihood parameters, first proposed by Dempster et al. (1977). The idea behind the finite mixture model for clustering is that the data originated from a mixture of subpopulations with each having different distributions. The process for finite mixture clustering involves randomly initialising parameters of different subgroups, evaluating how the observed data support the current distributions specified by the given parameter, and updating the parameter iteratively until the model converges (Fig. 6). When the observed variables follow normal distributions, the finite mixture model becomes a Gaussian mixture

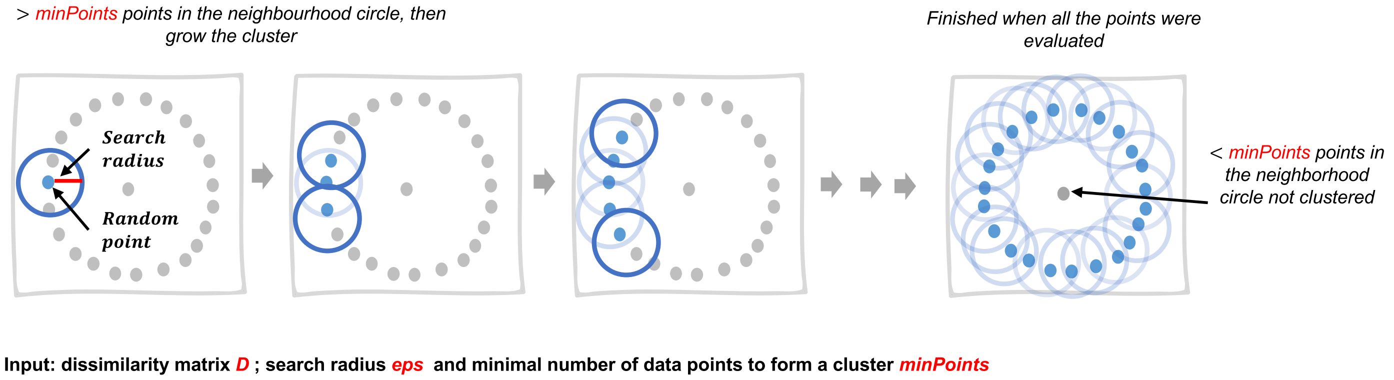

> minPoints points in the neighbourhood circle, then grow the cluster
Search radius
Random point
Finished when all the points were evaluated
< minPoints points in the neighborhood circle not clustered
Input: dissimilarity matrix D ; search radius eps and minimal number of data points to form a cluster minPoints

Fig. 5. Illustration of DBSCAN
Input: dissimilarity matrix D; search radius eps and minimal number of data points to form a cluster minPoints
Note: The concept of DBSCAN clustering is similar to disease transmission models, which involves the following steps: (i) randomly identify an initial point; (ii) use a radius (eps, user-defined) to find its neighbours; (iii) if there are enough neighbours (over minPoints, user-defined), the point will be considered as a core point, and its neighbours will be included in the current cluster, if not this point will stop "infecting" other points; (iv) iteratively repeat (ii) and (iii) for the newly classified points to establish the "infection" chain of the current cluster; (v) then select a different point that hasn't been "infected" as the initial point for a new cluster. The algorithm finishes when all the points have been evaluated. The points that are located in the lower density areas (without enough neighbours to establish a separate cluster) will be left un-clustered. The minPoints is normally selected based on domain knowledge or twice the dimensions of the data (number of variables), and eps can be specified by evaluating distance to the (2 × data dimensions +1) nearest neighbor (Sander et al., 1998; Schubert et al., 2017).

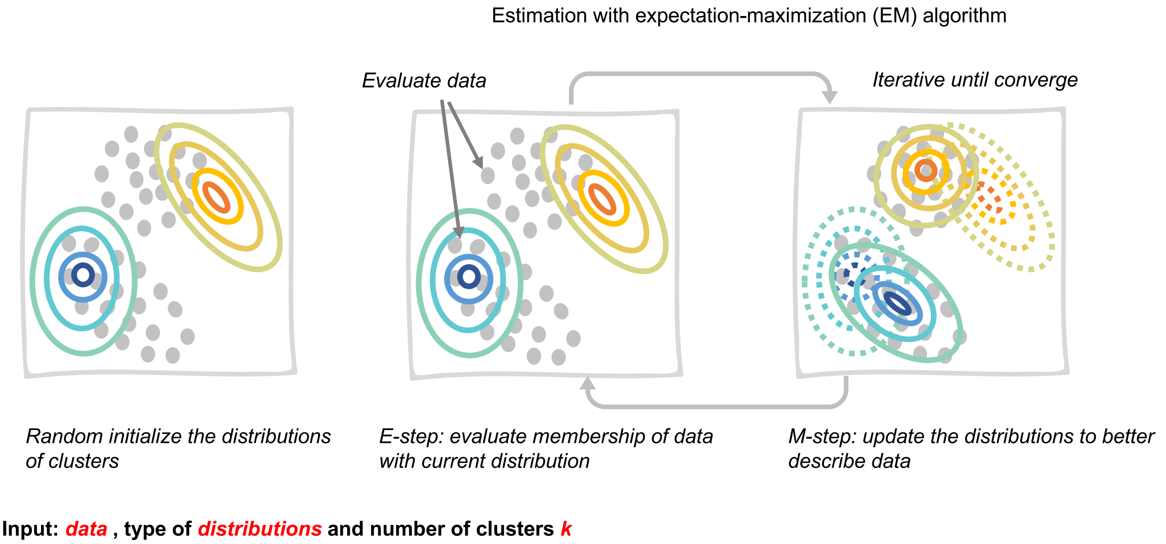

Estimation with expectation-maximization (EM) algorithm
Evaluate data
Iterative until converge
Random initialize the distributions of clusters
E-step: evaluate membership of data with current distribution
M-step: update the distributions to better describe data
Input: data , type of distributions and number of clusters k

Fig. 6. Illustration of Finite Mixture Model.
Note: Finite mixture models require specifications of the type of distribution for each variable and the number of clusters. The programme will: (i) first randomly assign parameters defining each cluster, and (ii) update modelling parameters using the EM algorithm until convergence. In the E-step each observation is assigned a weight for each cluster, and in the M-step the modelling parameters will be updated according to the weights estimated in the E-step.

model, and this type of clustering method is known as the latent profile analysis (LPA). When observed variables are discrete, the finite mixture model clustering is also known as the latent class analysis (LCA, Oberski 2016). In practice, the finite mixture model can also work with variables with mixed distributions (Wallace and Dowe, 2000). Model-based clustering algorithms are popular in mental health research due to their similarity in concept as well as model fitting routine with other latent variable modelling methods (e.g., structure equation modelling and factor analysis). Model-based algorithms can also directly estimate parameters (e.g., item response probabilities for LCA) that can facilitate interoperating models (e.g., how individual variables are associated with latent subgroups).

The R implementation methods, advantages and disadvantages for most of these common clustering algorithms are provided in Table 3.

2.3. *Extensions of common clustering algorithms*

Developing new clustering methods has been one of the key machine learning areas with extensive research work in the past decades. A range of advanced models have been developed to address different types of issues, for example, non-linear cluster boundaries, high dimension and sparse data, complex data structure, improvement of stability, special data types, and scalability for big data. Here we briefly introduce some of the state-of-art applications to facilitate better applications of these models.

2.3.1. *Kernel method*

One major limitation of many clustering methods is their difficulty in dealing with non-linear boundaries between clusters. An efficient method to deal with this issue is to use the kernel method first introduced by Aizerman (1964). The fundamental idea of the kernel method

Table 3
Types of Clustering Algorithms and Implementations in R.

Algorithms	Notes	R functions	References
Center-based clustering			
K-means	K-means requires an input of the Euclidean distance matrix. However, some studies suggest that other distance measures, such as Manhattan distance works better in a high dimensional space (Aggarwal et al., 2001). *Advantages*: fast to estimate and easy to interpret *Disadvantages*: assumes clusters have spherical shapes; sensitive to noise and outliers; global optimum is not ensured therefore sensitive to starting values	stats::kmeans cclust::cclust	Hartigan and Wong (1979), MacQueen (1967), Forgy (1965), Lloyd (1982)
K-means++	An algorithm that optimises the initialisation of k-means *Advantages*: not sensitive to starting values *Disadvantages*: same as k-means; requires slightly longer estimation time compared with K-means	pracma::kmeanspp	Arthur and Vassilvitskii (2006)
PAM (Partitioning Around Medoids)	A k-medoids algorithm uses a greedy search method, which, although it may not find the optimal solution, is faster compared with using an exhaustive search. *Advantages*: allows any distance matrix; more robust to outliers compared with K-means *Disadvantages:* slower to estimate with large dataset; does not work well with non-spherical clusters	cluster::pam fastkmedoids::fastpam A faster implementation of PAM using C++.	Kaufman and Rousseeuw (1990) Schubert and Rousseeuw (2019)
CLARA (Clustering Large Applications)	Extension of PAM for larger datasets using a randomly sampled smaller dataset. *Advantages*: faster than PAM *Disadvantages:* same as PAM; if the sample is biased, the best medoids cannot be guaranteed	cluster::clara fastkmedoids::fastclara	Kaufman and Rousseeuw (1990) Schubert and Rousseeuw (2019)
CLARANS (Clustering Large Applications based on Randomized Search)	Extension for CLARA. CLARANS does not draw a fixed sample of the data set at the beginning of the search, instead, it draws a new sample of neighbours in each step of a search. *Advantages*: faster than PAM; full data better represented compared with CLARA *Disadvantages:* same as PAM; sensitive to sequence of input data	fastkmedoids:: fastclarans A faster implementation of CLARANS using C++.	Ng and Han (2002)
K-medians	Use Manhattan distance matrix. *Advantages*: more robust to outliers compared with k-means *Disadvantages*: can only identify multidimensional spherical clusters; global optimum is not ensured therefore sensitive to starting values	Gmedian::KGmedian cclust::cclust	Cardot et al. (2012) MacQueen (1967)
K-modes	K-means style method for categorical variable. The model uses simple matching distance. *Advantages*: efficient for categorical variable; fast to estimate *Disadvantages:* cannot work with input data with a mixture of categorical and continuous variables; as the simple matching distance is used, the method does not take into account the "double-zero problem"; global optimum is not ensured therefore sensitive to starting values	klaR::kmodes	Huang (1997b) .
K-prototypes	K-prototypes is an integration of K-means and K-modes clustering. It employs different weights for continuous and categorical variables to avoid favouring any. *Advantages*: similar to K-means and K-modes *Disadvantages:* similar to K-means and K-modes; results can be sensitive to the weighting parameter	clustMixType::kproto	Huang (1997a)
Fuzzy C-means	Soft (fuzzy) extension of k-means clustering. *Advantages*: can identify overlapping clusters; robust to outliers *Disadvantages:* assumes clusters have spherical shapes; requires longer estimation time compared with K-means	e1071::cmeans	Bezdek (1981)
Kernel K-means	Kernel extension of K-means clustering. *Advantages*: can identify non-spherical cluster. *Disadvantages:* need specification of kernel function; slower compared with k-means, particularly with increasing observations and dimensionality; standard kernel K-means has a bias towards small high-density clusters (Marin et al., 2019)	klic::kkmeans kernlab: kkmeans weighted kernel K-means	Gönen and Margolin (2014) Inderjit S Dhillon et al. (2004)
Hierarchical clustering			
Agglomerative Hierarchical Clustering	Agglomerative Hierarchical Clustering is a Bottom-up clustering method, which is good at identifying small clusters.Hierarchical clustering using single linkage and complete linkage can cluster nonelliptical clusters, but are very sensitive to outliers and variation in density. Other methods are more robust to outliers, particularly the Ward's methods (Murtagh & Contreras, 2012), however do not work well with non-spherical clusters. *Advantages*: can work with any distance matrix cluster; all number of clusters estimated at the same time; dendrogram provides direct interpretation *Disadvantages:* mistakes made early in tree building stages cannot be fixed further down; slow to estimate with large dataset.	stats::hclust cluster::agnes Fastcluster:: hclust Note: C++ library for fast implementation of hierarchical agglomerative clustering.	Kaufman and Rousseeuw (1990) Müllner (2013)
Divisive hierarchical clustering (DIANA)	Top-down clustering, good at identifying large clusters. *Advantages*: same as agglomerative hierarchical clustering *Disadvantages:* same as agglomerative hierarchical clustering with only one criterion for dividing clusters (maximum average dissimilarity that is similar to average linkage in agglomerative hierarchical clustering); slower to estimate compared with agglomerative hierarchical clustering	cluster::diana	Kaufman and Rousseeuw (1990)

(*continued on next page*)

Table 3 (*continued*)

Algorithms	Notes	R functions	References
Hierarchical k-means clustering	Hybrid approach using hierarchical clustering results as the starting point for k-means clustering. *Advantages:* not sensitive to random seed *Disadvantages:* slower to run and lack of scalability compared with k-means	factoextra:: hkmeans	Milligan (1980)
CURE (Clustering Using REpresentatives)	A hierarchical agglomerative clustering operates on the representative points instead of individual data points. *Advantages*: more robust to outliers; can identify non-spherical clusters. *Disadvantages:* more time consuming to estimate on large datasets; sampling can be used to improve speed; comparable with single linkage method, therefore does not consider aggregated interconnectivity between clusters (Karypis et al., 1999)	Implementation only available in Python	Guha et al. (1998)
BIRCH (Balanced Iterative Reducing and clustering Using Hierarchies)	Multi-level clustering approach, which involves first building a clustering feature tree to obtain micro-level subclusters and then applying hierarchical clustering on established subclusters to obtain macro-level clustering results. *Advantages*: fast to implement; robust to outliers; can be used with other clustering methods *Disadvantages:* does not work well with non-spherical clusters; cannot work with non-numerical variables; sensitive to the order of the data record	Implementation only available in Python	Zhang et al. (1996),
Chameleon	Hierarchical clustering using dynamic modelling, which first establishes subclusters using k-nearest-neighbour graph, and then applies hierarchical agglomerative clustering to merge subclusters. *Advantages*: good at identifying arbitrarily shaped clusters; more robust to outliers and density variations. *Disadvantages:* more time consuming to estimate on large dataset; does not work well for high dimensional data	Implementation only available in Python	Karypis et al. (1999)
ROCK (Robust Clustering Using Links)	Hierarchical clustering designed for clustering categorical data (using Jaccard distance). It implements sampling and clustering using similarity graph (Linkage of points with common neighbours exceeding chosen distance threshold). *Advantages*: theoretically more appropriate for categorical data *Disadvantages:* random sample can be a biased presentation of the full dataset; more time consuming; threshold parameter is difficult to define	cba::rockCluster	Guha et al. (2000)
Density-based clustering			
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)	Clustering based on density distribution. *Advantages*: robust to outliers; can cluster arbitrarily shaped clusters; best number of clusters automatically estimated *Disadvantages:* sensitive to density variations; does not work well in high dimensional space; performance and quality of clusters highly dependant on parameter settings (e.g., running time increases with increasing neighbour radius, eps)	fpc::dbscan dbscan::dbscan Note: This implementation is faster and can work with larger data sets than dbscan in fpc.	Ester et al. (1996b).
OPTICS (Ordering Points To Identify the Clustering Structure)	The algorithm orders data points so that points with closer distance become neighbours. It also measures reachability distance between points which has a density interpretation: if a point is located in a low density area the reachability distance will be the distance to its nearest neighbours; if a point is located in a high density area the reachability distance will be the distance between two points. The order and reachability distances will allow the algorithm separate clusters with different levels of density in a hierarchical structure. *Advantages*: robust to outliers; can cluster arbitrarily shaped clusters; not sensitive to density variations; less sensitive to parameter settings; can extract clusters hierarchically *Disadvantages:* cannot select the best number of clusters automatically; does not work well in high dimensional space	dbscan::optics	Ankerst et al. (1999)
CLIQUE (Clustering in QUEst)	CLIQUE combines the design of density-based, grid-based and subspace clustering. It tries to: discretise data in individual dimensions so that high density regions can be identified; identify subspaces that may contain information on how clusters can be separated; and cluster using information on whether the dense units were connected in identified subspaces. *Advantages*: works well with high dimensional data; can identify arbitrarily shaped clusters; not sensitive to sequence of input data *Disadvantages:* quality of cluster highly depend on input parameters	Subspace::CLIQUE.	Agrawal et al. (1998),
HDBSCAN (Hierarchical DBSCAN)	Designed with a concept similar to OPTICS, quantifies distance between points with respect to their local density (using mutual reachability distance, which is similar to reachability distance in CLIQUE). Based on the distance matrix, the algorithm then estimates the minimum spanning tree (a hierarchal tree which connects all points in the minimal way), which could be simplified into a cluster dendrogram. The ingenious design of the algorithm makes it equivalent to running many DBSCANs with different radius parameter settings. *Advantages*: robust to outliers and density variations; can identify arbitrarily shaped clusters; only one parameter is needed in the model *Disadvantages:* may classify points as outliers unnecessarily	dbscan::hdbscan	Campello et al. (2013)
DPC (Density Peak Clustering)	DPC assumes that cluster centers should have higher densities and be relatively far apart. The algorithm estimates two properties: (1) local density of points (number of points within d_c distance) (2) minimal distance from the point to any other points with a higher density. Therefore, the cluster centers can be identified as those with higher local density but larger distance to other points with higher density (potential cluster centers). The remaining points are then assigned to the same cluster as its nearest	densityClust:: findClusters	Rodriguez and Laio (2014)

(*continued on next page*)

Table 3 (*continued*)

Algorithms	Notes	R functions	References
	neighbour of higher density. *Advantages*: cluster is assigned in a single step without iterative optimization, so very fast to converge; can identify arbitrarily shaped clusters; robust to noise *Disadvantages:* the cut-off parameter d_c can be difficult to set; cluster centers need to be selected manually; sensitive to density variations; can have errors when data are evenly distributed within clusters (Li & Zhang, 2020)		
Model-based clustering			
Finite mixture model	Can be used to fit a range of models and involves latent structures with a mixture of distributions, such as latent class analysis, latent profile analysis. *Advantages*: relatively robust to outliers and density variations; can identify elliptical and overlapping clusters; can cluster data with a mixture of distributions *Disadvantages:* cannot identify arbitrarily shaped clusters; assumption of local dependence can be easily violated; model may converge to a local maximum solution instead of a global maximum solution	depmixS4::mix Note: depmixS4 can also fit hidden Markov models. Mixtools::regmixEM flexmix::stepFlexmix Note: flexmix is highly flexible with self-defined M-step. Stochastic EM can be used to improve convergence to local optimum issue (Grün and Leisch, 2008).	McCutcheon (1987), Visser and Speekenbrink (2010) Benaglia et al. (2009) Friedrich (2004)
Latent profile analysis (LPA)/ Gaussian mixture model clustering	Gaussian mixture model is a special form of finite mixture model, when it is used for clustering it is also known as the latent profile analysis. *Advantages*: similar with finite mixture model; prior assumptions can be made about the shape (spherical or ellipsoidal), volume, and orientation of clusters (Scrucca et al., 2016) *Disadvantages:* similar to finite mixture model; can only cluster continuous variable	mclust::Mclust mclust::MclustBIC	Scrucca et al., (2016)
Latent class analysis (LCA)	LCA is a special type of finite mixture model for clustering with discrete variables. *Advantages*: similar to finite mixture model *Disadvantages:* similar to finite mixture model; may require a large number of parameters to be estimated (many variables with multiple categories); cannot model the ordinal data structure directly except when restricted LCA is used (Croon, 1990), which, however, can impose unrealistic constraints and can be difficult to fit; regularized models were suggested to provide promising results (Robitzsch, 2020).	poLCA::poLCA	Bandeen-roche et al. (1997)

is rather simple (Fig. 7). As non-linear data is difficult to be linearly separated in the original feature space, it aims to project the data into a higher dimensional feature space via a nonlinear mapping (e.g., from x to $x,\ x^2$). The ingenious design of the kernel method is that rather than explicitly working with the mapping functions (substantially increasing modelling parameters), it works with the dot product of the mapping function to save computational cost. The kernel method has been widely used in supervised learning methods such as support vector machines (SVM), and it can also be used in clustering algorithms to deal with nonlinear boundaries (Filippone et al., 2008). The most well-known kernel clustering method is the kernel K-means (Dhillon et al., 2004). Isolation kernel can also be used to define similarity (follows the same principles in human-judged similarity: e.g., Caucasians are less similar when compared in Europe than in Asia), which can improve DBSCAN's sensitivity to density variation issue (Qin et al., 2019).

2.3.2. *Neural network and deep-clustering*

Over the past few years, neural network and deep-learning

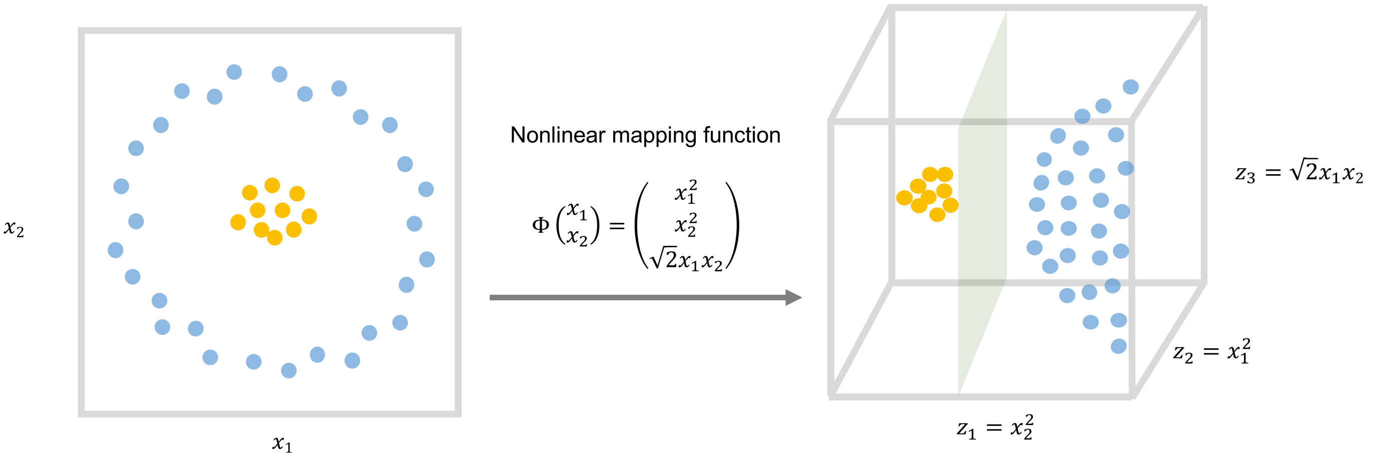

x_2
x_1
Nonlinear mapping function
$\Phi\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}x_1^2\\x_2^2\\\sqrt{2}x_1x_2\end{pmatrix}$
$z_3=\sqrt{2}x_1x_2$
$z_2=x_1^2$
$z_1=x_2^2$

Fig. 7. Illustration of Concept of Kernel Method.

Note: The kernel method transforms the data into higher-dimensional data via a nonlinear mapping function. The algorithm directly works with a kernel function $\mathrm{k}(\mathrm{x},\mathrm{x}')=\Phi(\mathrm{x})^{\mathrm{T}}\Phi(\mathrm{x}')$ (dot product of the mapping function) without having to work with the mapping function $\Phi(\mathrm{x})$. Popular kernel functions include linear kernels, polynomial kernels, and gaussian kernels.

algorithms have been making major advances in solving prediction and other machine learning problems in many applied areas such as image recognition, natural language processing, genomics, neural science and medicine (LeCun et al., 2015). An important application of deep learning is to learn and extract compressed and rich data features (also referred to as deep representations) from high dimensional data (Goodfellow et al., 2016). Deep representations can capture non-linear, additive and multiplicative effects in the data in lower dimensions. This makes deep learning methods a natural extension to be used in clustering algorithms.

The simplest method to apply a deep clustering algorithm is to extract deep representations of the data using deep learning methods first (which can be thought of as a dimensionality reduction technique) and then apply the common clustering algorithm based on the extracted features instead of the raw data. The simplest and most widely used approach is the autoencoder (Vincent et al., 2008), a type of neural network that compresses high dimensional data with complex structure (e.g., with non-linearity and interactions between variables) into low dimensional space via data reconstruction (Fig. 8). A range of other neural network or deep learning methods, such as Restricted Boltzmann Machines, variational autoencoder, convolutional neural network and recurrent neural network, can all be used in a similar encoder and decoder process to learn deep representations, see details discussed by Dara and Tumma (2018) and Min et al. (2018). Another alternative method to extract data representation for clustering is using Self-Organizing Map (SOM), a neural network that extracts data features on a two-dimensional space using hidden neurons (Lampinen and Oja, 1992). More recent developments in this area allow joint processes of deep learning and clustering, via optimizing a combined/joint loss function from both of them in an iterative process between them to improve model performance or a self-supervised learning framework (Zhang et al., 2019). A detailed introduction to these methods can be found elsewhere (Karim et al., 2020; Nutakki et al., 2019).

2.3.3. Semi-supervised clustering method

Semi-supervised clustering is a term that broadly refers to clustering based on partially available labelling information. It is widely used when clustering results are known for only a fraction of the data or when there is underlying known constraints in the data structure, see a summary of the recent development in these methods by Qin et al. (2019). Another type of semi-supervised clustering method, with a great level of relevancy in mental health, is a group of clustering algorithms that deal with a known feature being a “noisy surrogate” or closely related to clusters (Bair, 2013; Chand et al., 2020). The benefits of these models are their ability to include additional data (e.g., healthy controls), and clustering based on associations between risk factors and the outcome in clinical research.

There are three types of conceptual models developed to take special consideration of the outcome variable (Fig. 9). The first type is to apply additional control of the algorithm using the known outcome. For example, applying a pre-screening procedure to select variables (e.g., a simple statistical test to select variables that showed association with clinical outcomes) (Bair and Tibshirani, 2004; Koestler et al., 2010). Instead of variable selection, weights for variables according to their association with the outcome can also be used, e.g., Supervised Sparse Clustering (SSC) (Gaynor and Bair, 2017).

Another type of these clustering algorithms relies on the by-product of prediction models of the outcome variable as a measure of similarity between data points, e.g., Functional Random Forest (FRF) (Feczko et al., 2018) and deep learning methods (Eberle et al., 2022; Girish et al., 2019; Mathisen et al., 2020). The third type involves directly separating subgroups when establishing the prediction model, e.g., Heterogeneity Through Discriminative Analysis (HYDRA) (Varol et al., 2017). These models can have great potential in understanding heterogeneity that explains different causal mechanisms of mental health disorders and identify associated risk factors.

2.3.4. Time-series clustering algorithms

Time-series clustering is a special type of clustering algorithm that deals with data with a temporal structure such as aggregated health service use counts, neuroimaging data (e.g., fMRI, magnetoencephalography and electroencephalography), and longitudinal follow-up data. Although it can be treated the same as cross-sectional data, ignoring the temporal pattern can introduce substantial issues in clustering. One common method to address this issue is to apply similarity measurements to capture the time dimension. The similarity measures can be based on measuring the shapes of time series (Fig. 10). For example, Dynamic Time Warping (DTW) (Berndt and Clifford, 1994) and Longest Common SubSequence (LCSS) (Vlachos et al., 2002) were designed to find optimal alignment between series that do not sync up perfectly in the time domain. The similarity measures can also be compression based or model-based, as summarized by Aghabozorgi et al. (2015).

Many of these methods work well with shorter series, however, when the time series is long (high-dimensional), they can become intractable and less efficient (Wang et al., 2006). In this case, extracting important features (e.g., trends, cycles, and levels of noise) of the time series is needed before clustering. Popular methods include the Discrete Wavelet Transform (DWT), the Discrete Fourier Transform (DFT), and Piecewise Linear Approximation (PLA) (Aghabozorgi et al., 2015). Common time series models such as Hidden Markov Model (HMM), Auto-Regressive Moving Average (ARMA) and time series decomposition can all be used for feature extraction for clustering (Aghabozorgi et al., 2015; Wang et al., 2006).

An alternative method to cluster time series is to directly model the data generation process (model-based). Many of these models (Fig. 11), such as Growth Mixture Modelling (GMM) and Latent Class Growth Analysis (LCGA), have been widely used in mental health research (Jung and Wickrama, 2008). However, these models do not work with trajectories with arbitrary shapes, as they all assume linear or quadratic trends in the time domain. These models can be extended via the broader HMM framework that models the latent groups that change with time, for example, Latent Transition Analysis (LTA) (Collins and Lanza, 2009) and Random Intercept Latent Transition Analysis (RI-LTA) (Muthén and Asparouhov, 2020). In these models, multivariate time series (multiple variables changing with time) can also be evaluated. Traditionally, these models, particularly GMM and LCGA, were commonly used to understand non-linear trajectories in the mental disorder progression (Cole et al., 2012; Reef et al., 2011). However, more complex data collected from ecological momentary assessment brought new challenges (e.g., high dimensionality, sparsity, periodicity and noise) for these traditional methods and promoted the increasing use of more flexible models (Booij et al., 2021).

2.3.5. Graph and network-based algorithms

Graphs (also known as networks) are powerful mathematical abstractions that can describe complex systems of relations and interactions in fields ranging from biology and high-energy physics to social science and economics. Although real-world network data such as social networks (Fiori et al., 2006) and brain networks (Sporns, 2018) are obvious modalities for modelling using graphs, similarity measures can also be constructed as graphs, for example, association networks (van Borkulo et al., 2015) and Nearest neighbor Graph (Eppstein et al., 1997). On the graphs, each node represents a data entity (e.g., an individual or a variable), and an edge represents a connection, which could be directional, un-directional, weighted or unweighted. Like the other clustering methods, the key of graph clustering is to define the similarity among the nodes, which can be based on the density or pattern of the graph (Fig. 12). Having defined the similarity measure, similar clustering algorithms described above can be applied. Spectral Clustering (von Luxburg, 2007) and SCAN (Xu et al., 2007) are common methods used for clustering graph data. In mental health research, graph theory and related clustering models are widely used in neuroscience (Farahani et al., 2019), however, graph-based clustering models can also

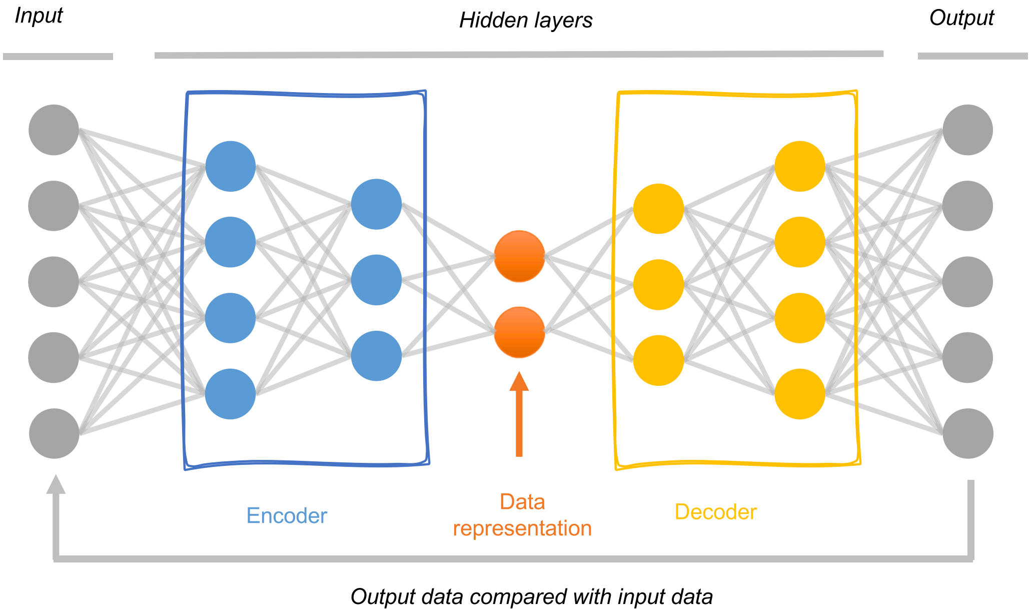


Fig. 8. Illustration of an Autoencoder.
Note: The autoencoder is a type of neural network that is used to learn a representation of the complex input data with an encoding process followed by a decoding process to reconstruct the data as close to the original input as possible. The encoder is a neural network that reduces input dimensions, whereas the decoder attempts to reconstruct the input data from the compressed dimensions. The autoencoder is trained (optimised) by minimizing the reconstruction error from the output back to the input. When there are multiple hidden layers in the encoding and decoding process, the autoencoder is known as the deep autoencoder.

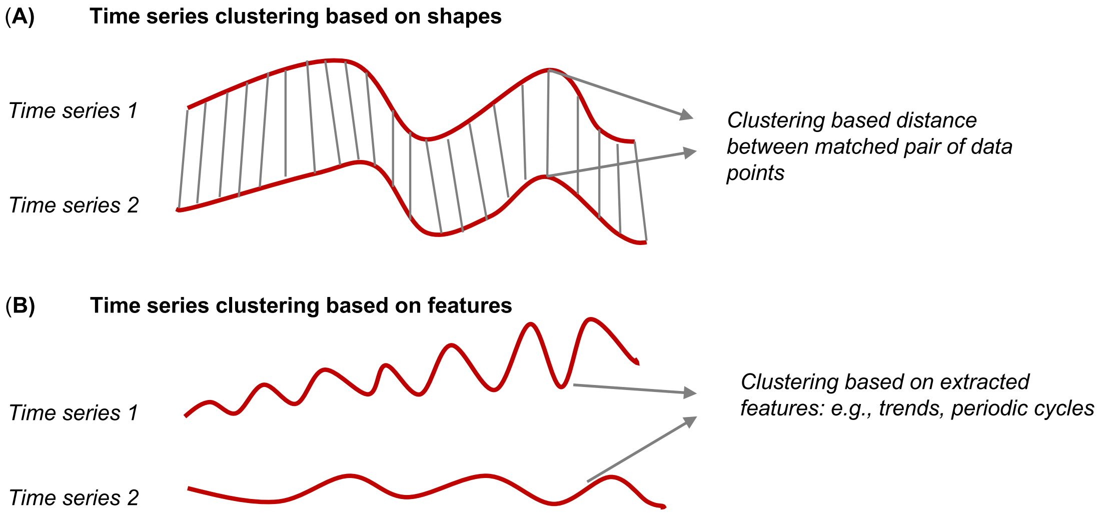

(A) Time series clustering based on shapes
Time series 1
Time series 2
Clustering based distance between matched pair of data points
(B) Time series clustering based on features
Time series 1
Time series 2
Clustering based on extracted features: e.g., trends, periodic cycles

Fig. 9. Semi-supervised Clustering Methods with Known Outcome.
Note: (A) Variables can initially be selected according to their association with the outcome variable and then be used in clustering algorithms. As this model can become problematic when variables are only weakly associated with the outcome, SCC was developed to give different weights to variables according to their associations with the outcome. (B) Prediction models of the outcome can produce by-products that can be interpreted as similarity measures between data points, e.g., proximity matrix (frequency of data points being classified into the same terminal node) estimated from random forest or similarity measures estimated from deep learning. (C) Specific models can directly identify subgroups within the same outcome group. HYDRA applied a Convex Polytope Classification, which can be thought of as combining multiple prediction models, with each trying to separate (linearly) between those with and without the outcome.

be used in evaluating symptom networks (Brusco et al., 2022) or modelling other complex data such as text (Preoţiuc-Pietro et al., 2015).

2.3.6. Bayesian clustering models

Unlike non-Bayesian models, Bayesian models apply Bayes' theorem, which uses observed data to update available knowledge (prior distribution, e.g., prior belief of treatment response) to obtain a more accurate understanding of the parameter (posterior distribution) (van de Schoot et al., 2021). In clustering algorithms, Bayesian methods provide the natural extension from hard clustering to soft clustering, see the soft K-means model in Stan (Stan Development Team, 2019). Another important application of the Bayesian method in clustering is in Finite Mixture Models, such as Dirichlet multinomial mixture model (Yin and Wang, 2014) and Gaussian mixture model (He et al., 2011; Manduchi et al., 2021), which provide more flexibility in model design and robustness in inference. Bayesian mixture models can also be infinite. Non-parametric Bayesian infinite mixture models, such as Dirichlet process mixture model (Li et al., 2019) and Indian Buffet Process mixture for overlapping clusters (Griffiths and Ghahramani, 2011), allow the data to decide the number of latent clusters (without the assumption of the number of clusters).

2.3.7. Clustering ensembles

Clustering ensembles is a method that compares and aggregates results from different models. As most of the existing models have different shortcomings and many are sensitive to the choice of parameters,

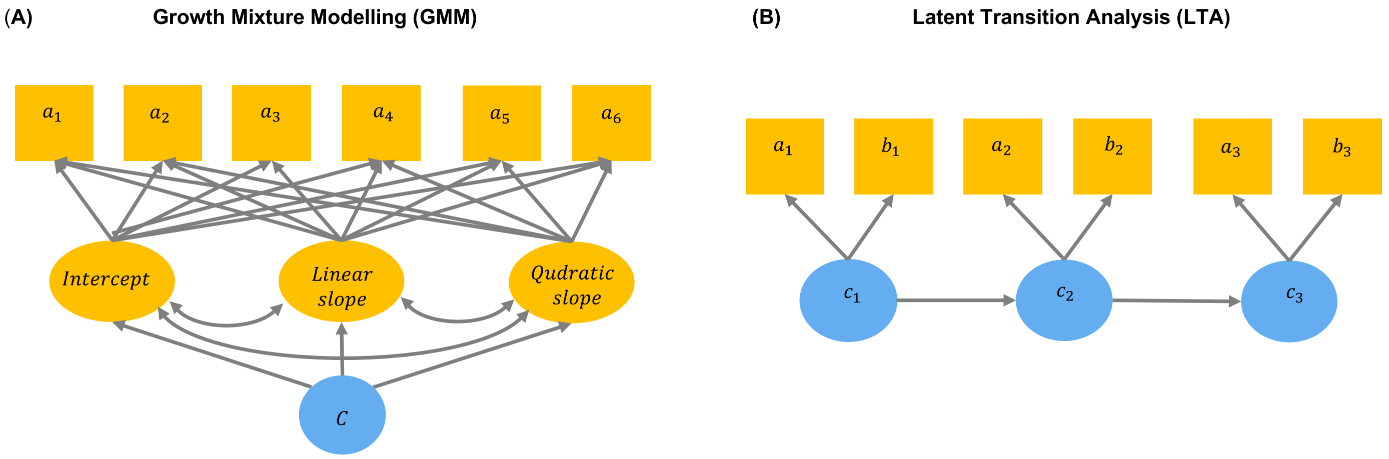


Fig. 10. Time Series Clustering.
Note: (A) Time series clustering based on similarity measures. As there are often lags present in the data, popular time series similarity measures try to first identify the best matching pairs of data points that may not sync up in the time domain. (B) Time series clustering based on features. When there are multiple series or long series, it is difficult to compare similar shapes. Therefore, features associated with individual series can be extracted to compute similarities.

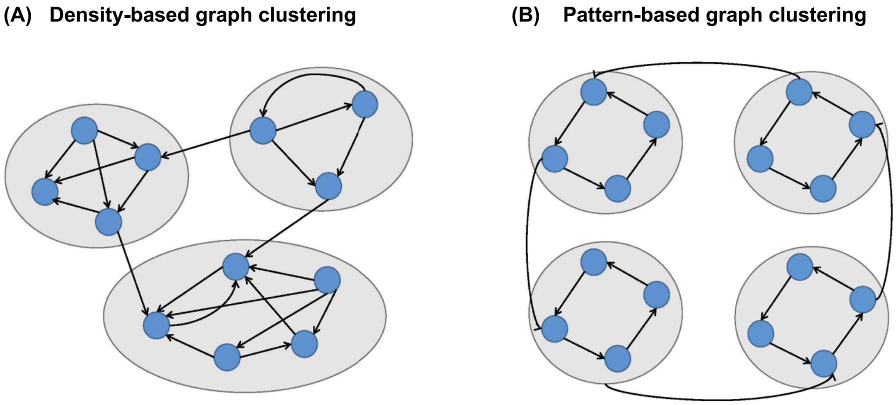

(A) Density-based graph clustering
(B) Pattern-based graph clustering

Fig. 11. Examples of Application of Mixed Model for Time-series Data.
Note: (A) Consider longitudinal observations of a_1, a_2, a_3, a_4, a_5, and a_6, GMM models latent subgroups c presenting with different intercepts and linear/nonlinear slopes. (B) LTA models the latent Markov process where observations at time points 1 (a_1, b_1), 2 (a_2, b_2), and 3 (a_3, b_3), are determined by their specific latent group and the latent group can transit between time points

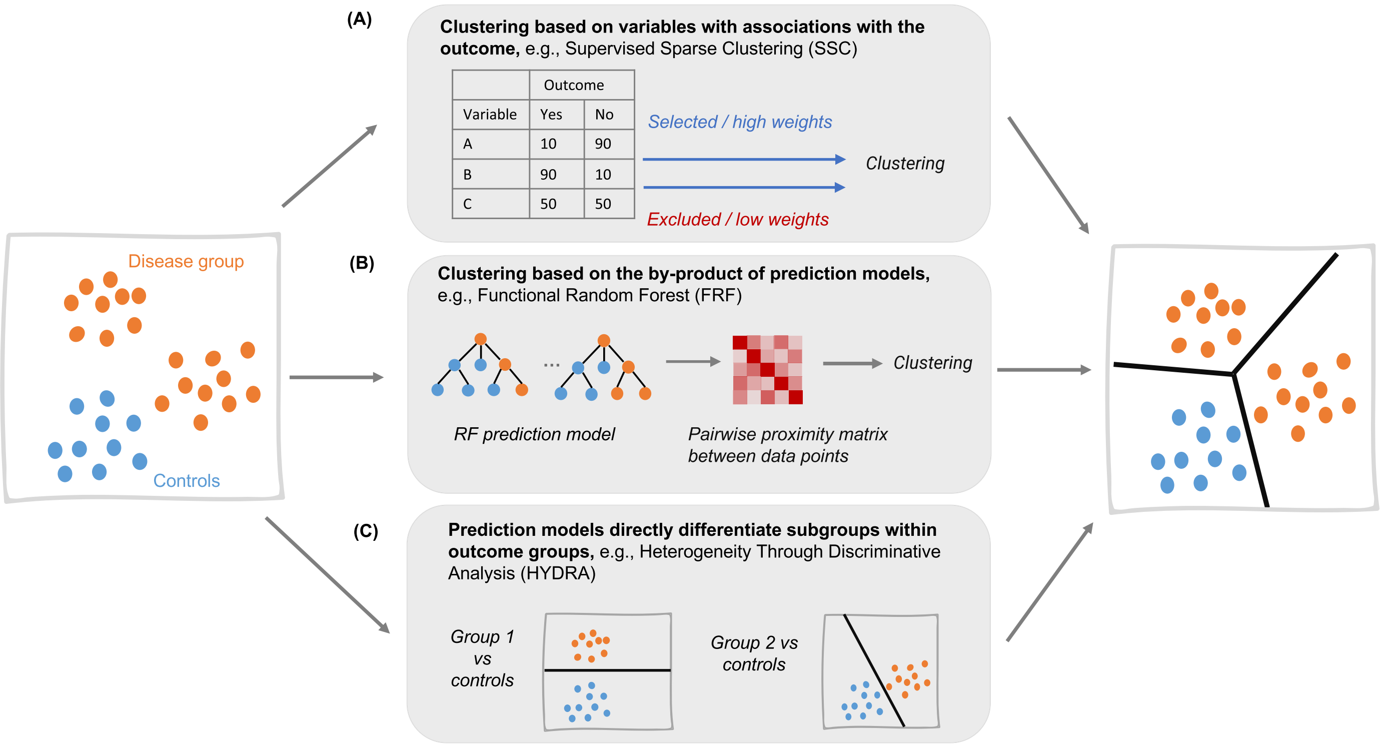
random seeds and minor data changes, pooling or aggregating multiple models can provide more robust and generalizable results. Clustering ensembles involves first running separate models, e.g., models using different methods (Chiu and Talhouk, 2018) or subsets of observations and/or variables (Dwyer et al., 2020; John et al., 2020), and then pooling the results using a consensus function, which can be defined in many ways (Vega-Pons and Ruiz-Shulcloper, 2011). Clustering ensemble methods were found able to improve the stability and robustness of results (Fred and Jain, 2002, 2005; Monti et al., 2003; Strehl and Ghosh, 2002; Topchy et al., 2004).

A commonly used consensus function in health science, particularly in genomic studies, is the co-association matrix (proportion of a pair of observations being clustered into the same cluster) (Monti et al., 2003). The pooled results can then be treated as similarity measures and be decomposed using other clustering methods such as hierarchical clustering or spectral clustering to obtain the final cluster membership. This method is commonly known as consensus clustering (although consensus clustering is sometimes used interchangeably with clustering ensembles). Comprehensive reviews on clustering ensembles can be found elsewhere (Boongoen and Iam-On, 2018; Vega-Pons and Ruiz-Shulcloper, 2011).

Clustering ensembles is one of the most promising methods towards high robustness , which offers flexible ways to explore the impact of randomness, clustering methods, parameters, and variables as well as aggregate results to reduce uncertainty. Ensemble methods had significant theoretical development over the past decade, however, these methods haven't been applied and evaluated extensively in practice. One notable suggestion by (Şenbabaoğlu et al., 2014) recommended that consensus clustering needs to be implemented with care in selecting the best number of clusters as partitioning may appear to be stable when subclusters do not exist (Şenbabaoğlu et al., 2014).

2.3.8. *Multiview clustering*

Another related extension, perhaps less familiar to mental health researchers, is clustering multi-view or multimodal data (data collected from multiple sources such as images, videos, audios, and biomarkers). Traditional clustering methods require complex data structures to be aggregated with single distance measures or modelled with single latent factors. Although advanced dimension reduction technics can be used, they can often miss important information related to different types of data sources and are biased towards data sources with more features (or higher variation in features). Multi-view clustering algorithms have, therefore, been developed to address these issues. This method takes advantage of naturally formed views of data representations (e.g., data


(A)
Clustering based on variables with associations with the outcome, e.g., Supervised Sparse Clustering (SSC)
Selected / high weights
Clustering
Excluded / low weights
(B)
Clustering based on the by-product of prediction models, e.g., Functional Random Forest (FRF)
Clustering
RF prediction model
Pairwise proximity matrix between data points
(C)
Prediction models directly differentiate subgroups within outcome groups, e.g., Heterogeneity Through Discriminative Analysis (HYDRA)
Group 1 vs controls
Group 2 vs controls
Disease group
Controls

	Outcome	
Variable	Yes	No
A	10	90
B	90	10
C	50	50

Fig. 12. Examples of Pattern-based Graph Clusters.
Note: (A) Examples of three density-based clusters. Graph clustering based on density-based similarity assumes that the nodes of the same cluster have dense connections while nodes across clusters have sparser connections. In this case, the input feature to the clustering algorithms is based on the graphs' edges where the number of edges between a set of nodes can be considered as a similarity node indicator in the graph. (B) depicts a graph with four pattern-based (flow) clusters. Pattern-based similarity measures include similar nodes that go beyond edge density patterns.

from different sources), and the complementary and consensus clustering information from individual views can then be summarized to create a final cluster solution (Chao et al., 2021; Fu et al., 2020; Yang and Wang, 2018) or multi-level solution (Dwyer et al., 2020).

2.3.9. *Other extensions*

Hundreds of new clustering algorithms are published yearly, and many of them are aimed at solving complex and real-world problems. One area of particular interest to mental health researchers is big data application. Traditionally, clustering models were applied to hundreds to thousands of records - in modern times, researchers may require clustering of thousands to millions of records (e.g., medical treatment data at the population level). In this case, some traditional models are not scalable to such large datasets. A range of algorithms have been developed to deal with this challenge, see methods summarized by Fahad et al. (2014).

3. Choosing algorithms and processes for different challenges

Most clustering algorithms work well when clusters are sphere-shaped and highly separable in low dimensions without outliers. However, in practice, the data collected is often high dimensional (e.g., neuroimaging data), noisy (e.g., administrative healthcare records), and complex (e.g., longitudinal studies). In this section, we summarize major challenges and how different models may be suitable for addressing these issues (Table 4).

3.1. *High dimensional, noisy and sparse data*

The majority of clustering algorithms have difficulties working with high dimensional data (e.g., over 100 variables) due to difficulties in defining efficient distance measures in high dimensional space (Steinbach et al., 2004). Model-based clustering algorithms do not require distance measures; however, they face issues such as difficulties in convergence and high computational costs. In some cases, there are more variables than the number of observations. Therefore, these models require variable selection, data aggregation, or dimensionality reduction when facing increasing data dimensions. Importantly, this process should be conducted robustly (e.g., using cross-validation [CV]) and meaningfully (taking clinical insights into account) to avoid overfitting.

There are also new methods available for working directly with high dimensional data (Steinbach et al., 2004). Subspace clustering, for example, searches and select variables best at separating potential clusters (Sim et al., 2013). It is useful when there are many irrelevant/noise features and a global dimensionality reduction becomes inefficient to identify the subset of features related to underlying clusters.

3.2. *Skewed distribution*

As center-based, model-based, and most hierarchical clustering algorithms require multivariate spherical clusters, substantially skewed distributions will cause issues. However, many clinical measures follow skewed distributions as distance measures also give higher weights to the tail of the distribution, causing bias in measurement. Data skewness also introduces density variations, which is problematic for density-based methods such as DBSCAN. Some kernel methods estimating local distributions/structures (Marin et al., 2017) and model-based clustering for skewed distributions can be used to reduce the impact of data skewness. However, these advanced methods can be difficult to implement, therefore, it is recommended to normalize skewed distributions before clustering. Data transformation was found particularly effective in reducing bias in the density estimators (Qin et al., 2019; Zhu

Table 4
Challenges in Clustering Tasks and Robustness for Different Models.

	Center-based clustering	Hierarchical clustering	Density-based clustering	Model-based clustering	Extensions
High dimensional, noise, and sparse data	Problematic	Problematic	Problematic	Problematic	Requires variable selection and dimensionality reduction or subspace clustering methods
Skewed distribution	Problematic	Problematic	Robust for models allowing density variation	Robust when distributions are correctly estimated	Data normalization is generally needed.
Outliers	Problematic	Problematic	Robust	Can be problematic	Fuzzy clustering is more robust to outliers. Outlier/anomaly detection models can be used prior to clustering.
Overlapping boundaries	Problematic	Problematic	Problematic	Robust	Fuzzy clustering can be used when there are overlapping cluster boundaries.
Arbitrary cluster shapes	Problematic	Likely to be problematic	Robust	Problematic	There are many extension models available such as kernel-based clustering and non-linear feature extraction models.
Rare events	Likely to be problematic	Likely to be problematic	Likely to be problematic	Likely to be problematic	Consider anomaly detection algorithms when aiming to identify very small clusters.
Mixed data	Potentially problematic	Potentially problematic	Potentially problematic	Potentially problematic	Distance measures for mixed data can be used, however can be sensitive to outliers and not capturing important features. Dimensionality reduction is needed prior to clustering.
Missing data	Problematic	Problematic	Problematic	Likely to be problematic	Multiple imputation models or clustering algorithms that specifically model data missingness are needed for data MAR.

et al., 2021).

3.3. Outliers

Outliers are extreme data values that differ significantly from other observations. Outliers are commonly depicted in univariate distributions; however, in clustering algorithms, outliers have to be evaluated on multivariate associations (points in the center of a univariate distribution can still be an outlier). Many commonly used algorithms such as K-means and hierarchical clustering are known to be sensitive to outliers (Zouridakis et al., 1997). Density-based, model-based and fuzzy clustering are more robust to outliers (Zouridakis et al., 1997). Outlier/anomaly detection models, such as LOF (Breunig et al., 2000) and iForest (Bandaragoda et al., 2018; Liu et al., 2008, Liu, 2010), can be used to identify outliers before clustering (Chandola et al., 2009).

3.4. Overlapping boundaries

In some cases, there may be overlap or ambiguity in underlying clusters (Chen et al., 2020). In this case, hard clustering methods can be problematic, and soft-clustering models, such as fuzzy clustering and model-based clustering should be used. Joint group membership can also be an independent research interest (Pantelis et al., 2003; Rovetta and Masulli, 2019). However, evaluating model performance and identifying appropriate group membership based on probabilities of group membership can sometimes be difficult for fuzzy clustering (Sato-Ilic and Jain, 2006).

3.5. Arbitrary cluster shapes

Most of the traditional partitioning-based and model-based clustering methods assume spherical, elliptical or convex shapes of clusters. Although single linkage hierarchical clustering can identify arbitrarily shaped clusters (Ros and Guillaume, 2019), it often fails to identify clusters in practice due to data noise and overlaps between clusters. More recent models, such as density-based, kernel-based and deep clustering, can work well with arbitrary cluster shapes and should be the models of choice if the boundaries between clusters are hypothesised to be non-convex.

3.6. Rare event

Imbalanced data can be challenging to work with for many machine learning algorithms, as they tend to be biased towards majority groups (Krawczyk, 2016). Small clusters can sometimes be very difficult to detect, but can often be important in clinical settings. The power of detecting small clusters mainly depends on how separable they are from main clusters, and the size of the cluster relative to the total sample size. Highly separable smaller clusters (e.g., two clusters with 9:1 ratio) are easy to identify using almost any method, but many existing methods cannot identify less-separable smaller clusters even with a large sample size (Dalmaijer et al., 2022). More advanced methods such as HDBSCAN and density peak clustering may provide better results. In some cases, finding smaller and less-separable clusters may, perhaps, be better considered as an anomaly detection task.

3.7. Mixed and multimodal data

In practice, researchers often need to deal with clustering tasks where input data is a mixture of different data types (e.g., continuous, nominal, binary, and ordinal) or from different data sources (e.g., survey, biomarker, and image). There are a few methods available to work with mixed data types. The easiest approach is to use distance measures for mixed data (e.g., Gower distance). However, this method can be sensitive to outliers and ignores important multivariate data features. An alternative approach is to apply a dimensionality reduction technique on mixed data, such as unimodal Variational Autoencoder (Simidjievski et al., 2019), Factor Analysis of Mixed Data (FAMD) (Pagès, 2014), or PCA for mixed data (PCAmix) (Chavent et al., 2014). An alternative regime is to operate in segmented datasets using methods such as subspace clustering and multi-view clustering. When data were obtained from combinations of psychological measures that can be summarized by underlying latent dimensions (e.g., from a combination of symptom severity measures, self-reported conditions), methods such as FAMD and PCAmix may be more attractive. However, when data were obtained from different sources with high dimensionality, deep learning, subspace clustering and multi-view clustering can be more useful to address the difficulties of linearly summarizing data.

3.8. Missing data

When data is Missing Completely at Random (MCAR), all the clustering models will be unbiased. However, MCAR is rarely the case in practice. The most common type of missing data is Missing at Random (MAR: missingness is related to observed data) or Missing Not at Random (MNAR: missingness is related to unobserved data). As most clustering methods cannot directly deal with missing data, multiple imputation (using methods such as multiple imputation using chained equations [MICE]) is commonly needed (Basagaña et al., 2013). A full information maximum likelihood model can be used for model-based clustering (Enders and Bandalos, 2001); however this model cannot take auxiliary variables (variables related to the missing data but not underlying cluster) into account. Although a few combined imputation and clustering models, such as k-POD (Chi et al., 2016) have been developed, the impact of missing data and imputation methods hasn't been evaluated extensively compared with inference and prediction models. A sensible and promising domain is to combine multiple imputation with ensemble clustering, i.e., ensemble results from multiple imputed datasets within the cross-validation framework (Chao et al., 2022; Pattanodom et al., 2016; Wan et al., 2020). Future studies are needed to establish the optimal pipelines for addressing missing data under different conditions.

3.9. *Statistical power*

Power estimation cannot be conducted for clustering analysis due to its exploratory nature. Whether clusters can be detected correctly depends on the sample size, number of parameters, how separatable clusters are, level of noise, relative size of clusters, and the method used (Dalmaijer et al., 2022; Dolnicar et al., 2013; Tueller and Lubke, 2010), all of which are largely unknown. Therefore, it is not feasible, or perhaps not appropriate, to determine the statistical power for clustering. Highly separable clusters without noise, though uncommon in practice, can be correctly detected with a small number of observations (Dalmaijer et al., 2022). However, researchers need to ensure that the modelling approach is feasible (e.g., not over parameterized, particularly for LCA), robust (e.g., sampling and CV provide consistent results), and generalizable (unbiased representation of the population of interest). A large number of clusters estimated from a small sample or clusters with very small sizes can be indicators of a lack of robustness and generalizability.

3.10. *Multilevel data*

Data can have a multilevel nature (e.g., students nested in schools, randomized cluster trial). Higher correlations may be observed in naturally formed clusters. In some cases, this multi-level feature is not problematic (e.g., participants recruited in one site may show more severe symptoms but do not differ in the latent heterogeneity of interest such as causal mechanisms of the disorder). However, sometimes it may bias clustering results (e.g., different neuroimaging machines used), or be a separate research interest. In these cases, multilevel mixture models can be used (Asparouhov and Muthén, 2008). Alternatively, data fusion models can be applied to obtain reliable and consistent information from raw data and remove the systematic noise (Meng et al., 2020).

3.11. *Poor stability*

All clustering algorithms can be impacted by the random starting point and minor data changes. The most widely known one is the local optima problem for K-means (Steinley, 2003). Essentially, the algorithm tends to obtain a suboptimal result (getting stuck in a local optimal solution, not a global optimal solution) when optimizing the loss criterion. Therefore, the exact shape of the input data, a change of random seeds, and any minor data perturbation, tend to generate different cluster memberships. Other methods largely suffer from the same issue (Goodman, 1974; Monti et al., 2003; van der Kloot et al., 2005). The lack of stability also impacts the choice of optimal modelling parameters such as the number of clusters. Although a few methods were developed to address this issue such as hybrid hierarchical k-means clustering (Milligan, 1980) and K-means++ (Arthur and Vassilvitskii, 2006), recently developed clustering ensemble methods offer a greater level of robustness when combined with resampling and CV.

3.12. *Clinically meaningful clusters*

Although there are many aspects to consider when applying a clustering algorithm in practice, one of the most important issues is to ensure whether the model can detect clinically meaningful clusters. In some cases (e.g., establishing distinct illness subtypes), it is meaningless to dichotomize a continuum (Dinga et al., 2019). In clinical practice, however, it is commonplace to have varying degrees of intervention along such a continuum (e.g., normal variant requiring no further follow-up, watchful waiting, brief treatment, invasive therapy) and separating groups with different severity may become useful (Cotton et al., 2022). What is critical is to have engaged meaningfully with both clinical teams and consumers in question to ensure that the outcome represents a target relevant to both parties.

4. Data pre-processing and testing

4.1. *Data pre-processing and dimensionality reduction*

In practice, data obtained for clustering are often high-dimensional, which introduces difficulties for most of the clustering algorithms. Therefore, an important stage in clustering tasks is data pre-processing, which often involves data normalization, variable selection and dimensionality reduction.

To facilitate the estimation of similarity measures or probability distributions, the input data need to be transformed to adjust in range, dispersion and skewness. The commonly used methods include z-score normalization, min-max normalization, quotient normalization and Box-Cox power transformation, see the detailed summary of proposed methods elsewhere (Jajuga and Walesiak, 2000; Milligan and Cooper, 1988). Double standardization (z-score normalization by columns and rows) can be used when only the relative differences between variables are associated with underlying clusters (similar scenarios to when a scale-invariant distance measure is preferred).

It is important to note that the type of normalization and transformation needed should be based on the clinical understanding of the data and how the variability, scale and shape of distribution may impact the difference between data points. For example, if the relative differences between individuals across variables are more important, z-score normalization is more suitable. However, when the raw score differences are more important, min-max normalization can be a better method as it preserves variability differences between variables.

As the data collected may not necessarily be related to underlying heterogeneity, including a high proportion of "useless" or "low quality" variables can often introduce additional difficulties for clustering algorithms. Therefore, it is important to identify and pre-select variables that have good data quality and are potentially related to heterogeneity or use advanced computational methods to deal with data noise and quality problems (Shen et al., 2014). Another issue, a substantial concern in mental health research rarely mentioned in clustering literature, is the need to avoid over-represented variables measuring the same construct. For example, if the researcher included nine individual items of PHQ-9 and the mean scores of GAD-7 in a K-means clustering, the distance measured between two participants would be highly reflective of their differences in depression but not in anxiety.

In practice, researchers are often required to further reduce data dimensions or suppress data non-linearity to ensure the efficiency of clustering algorithms. This process is known as dimensionality reduction which involves projecting the high dimensional space into a low dimensional space via a series of numerical operations based on the input data. The most commonly used dimensionality reduction techniques in psychology are principal component analysis (PCA) and factor analysis, which are both linear dimension reduction methods (Fodor, 2002; Jolliffe, 2022). In modern machine learning, a range of non-linear models such as kernel PCA, Non-negative matrix factorization (NMF), Graph Embedding, and autoencoder (discussed above) are widely used.

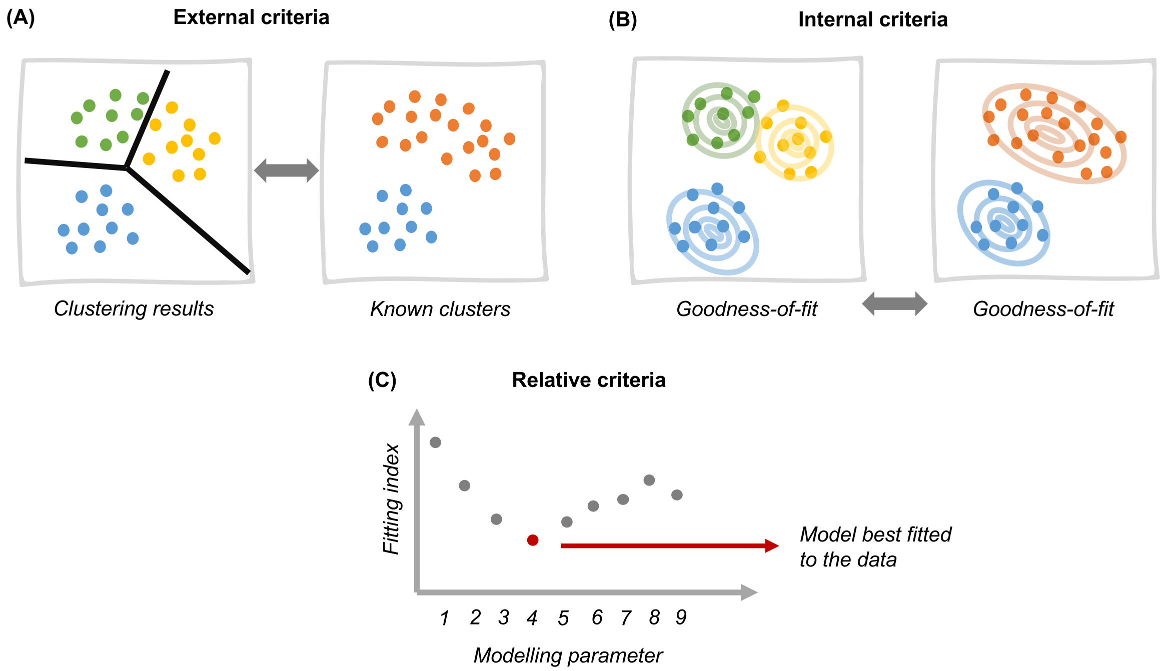


Fig. 13. Clusters Evaluation Criteria.
Notes: (A) External criteria commonly compare clustering results with an external reference (normally the 'ground truth' or 'gold standard' clustering results). These criteria are commonly used to theoretically compare models using benchmark datasets. (B) Internal criteria utilize information estimated as a part of the clustering process to determine the fit of the model to the data. (C) Relative criteria focus on evaluating how well different models represent the underlying data structure. Sometimes relative criteria are also classified as internal criteria as it does not involve external known information about group membership.

These methods can be applied more regularly in mental health research to ensure that nonlinearity and interactions amongst variables are not ignored. Comprehensive reviews of these methods are available (Cunningham and Ghahramani, 2015; Van Der Maaten et al., 2009) to guide applications of these methods.

4.2. Pre-clustering testing

A range of methods were proposed to test for the presence of multiple clusters in the data, see details summarized by Adolfsson et al. (2019). The basic design concept of the proposed clustering tests was to evaluate data multimodality (more than one mode indicates heterogeneity) or randomness (similarity with randomly generated data indicates lack of heterogeneity).

The most widely used and robust method is the Dip test on pairwise distances (Dip-dist), which calculates and sorts all pairwise distances between any two data points, and evaluates whether there is any "dip" in the continuous distribution of the pairwise distances (Hartigan and Hartigan, 1985). The Silverman test can also be used to test multimodality, which tries to find out how much smoothing (larger bandwidth for the kernel density estimate) is needed for the data distribution to approximate a normal distribution (Silverman, 1981). Both Silverman test and Dip test can be used to test multimodality on the Principal Component and Principal Curve (Adolfsson et al., 2019). To evaluate data randomness, the Hopkins test compares the observed data with randomly generated data under a uniform distribution. If the observed data do not have any clusters, the distances between a sample of data points with their nearest neighbors will be the same as the distances between a sample of simulated data points with their nearest neighbors (Lawson and Jurs, 1990).

It should be noted that different tests provide slightly different results depending on factors such as outliers, overlapping clusters, and non-linear boundaries (Adolfsson et al., 2019). Therefore, pre-clustering testing should be used as a guide rather than a hypothesis testing tool. Alternatively, clustering results and evaluation criteria from the data can be compared with synthetic data without subgroups, e.g., permutated or generated from a uniform or unimodal distribution (Gordon, 1996).

5. Cluster evaluation and cross-validation

5.1. Cluster evaluation

Due to the unsupervised nature of clustering algorithms, there is no consensus on many modelling choices such as the optimal distance measures, number of clusters, parameters and/or modelling technique. The appropriateness of modelling choices depends on the nature of the data and the underlying heterogeneity. As a result, it is common practice to employ a range of different methods and parameter choices to analyze a dataset and then conduct evaluation and validation.

To date, many methods have been developed to validate clustering results (Fig. 13), which can be broadly classified into external criteria (evaluate how well the model describes the truth, i.e., known subgroups), internal criteria (evaluate how well the model describes the data) or relative criteria (evaluate which model or modelling parameter produces best quality clusters) (Gan et al., 2020). A detailed summary of these criteria as well as R implementations are provided in Tables 5 and 6.

5.1.1. External criteria

The external criteria compare clustering results with the known underlying cluster structure. They are generally used with a theoretical framework to validate and compare different methods. They can also be used to estimate the stability of cluster assignments over resampled data. When real and predicted clusters are assumed to have a one-to-one mapping, methods such as F-measure can be applied (Amigó et al., 2009). Alternatively, the agreement can be measured by counting all pairs of data points and evaluating whether they were grouped into the same or different groups in real and predicted clusters (e.g., Rand statistic, Jaccard coefficient and Fowlkes and Mallows index [FM]). The

Table 5
External Validation Indexes and Implementation in R.

Indexes	Equation *k :number of clusters* *n: number of data points* *L:* real clusters *C*: predicted clusters	Note	R-packages	Refs.
F-measure [*]	$\frac{1}{k}\sum_{i=1}^{k}\frac{2\ P_i\ \times R_i}{P_i+R_i}$ $P_i\ =\frac{TP_i}{TP_i+FP_i}$ $R_i\ =\frac{TP_i}{TP_i+FN_i}$	Evaluate average combination of precision (P_i) and recall (R_i) for each identified predicted cluster relative to the matching real cluster.	FlowSOM:: FMeasure	Zaki et al. (2014)
BCubed F-score	$\frac{2\ P\ \times R}{P+R}$ $P\ =\frac{1}{n}\sum_{j=1}^{n}\frac{No.\ in\ same\ C}{No.\ in\ L}$ $R\ =\frac{1}{n}\sum_{j=1}^{n}\frac{No.\ in\ same\ L}{No.\ in\ C}$	Similar with the F-measures except for using the BCubed precision and recall for individual observations.	DPBBM:: BCubed_metric	Bagga and Baldwin (1998)
Rand index [ˆ]	$\frac{TP+TN}{TP+TN+FP+FN}$	Measuring similarity between clustering results and known clusters via counting pairs of data points.	aricode::RI clusteval:: rand_indep	Rand (1971)
Adjusted Rand index	$\frac{index\ \ -E(\ index)}{\max(index)-E(\ index)}$ $index\ =\sum_{i,j}\binom{n_{ij}}{2}$ $E(\ index)\ =\left[\sum_i\binom{C_i}{2}\sum_j\binom{L_j}{2}\right]\Big/\binom{n_{ij}}{2}$ $\max(index)\ =\left[\sum_i\binom{C_i}{2}+\ \sum_j\binom{L_j}{2}\right]\Big/2$	The corrected-for-chance version of the Rand index. It establishes the lower bound of 0 when the index is the same as the expected index, $E(\ index)$, which is the index from two completely random partitions.	aricode::ARI	Hubert and Arabie (1985)
Jaccard coefficient [ˆ]	$\frac{TP}{TP+FP+FN}$	Measuring similarly by excluding pairs of data points belonging to different groups in both real and predicted clusters (TN).	clusteval:: jaccard_indep	Halkidi et al. (2001)
Fowlkes and Mallows index (FM) [ˆ]	$\sqrt{P\times R}$ $P\ =\frac{TP}{TP+FP}$ $R\ =\frac{TP}{TP+FN}$	Measuring similarly of pairs of data points using the product of precision (P) and recall (R).	dendextend:: FM_index	Fowlkes and Mallows (1983), Halkidi et al. (2001)
Normalized mutual information (NMI) [§]	$\frac{I(L,C)}{\sqrt{H(L)H(C)}}$ $H(.)$: entropy $I(L,C)$: mutual information $I(L,C)\ =H(L)-H(YL\|C)$	Indicate the reduction in the entropy of real clusters if the predicted clusters are known. Higher NMI indicates better clustering results.	aricode::NMI	Vinh et al. (2009)
Adjusted mutual information (AMI) [§]	$\frac{I(L,C)-E(I(L,C))}{\max(H(L),H(C))-E(I(L,C))}$ $E(I(L,C))$: expected mutual information between two random clusters	Similar to NMI and corrects the effect of agreement solely due to chance between clusters.	aricode::AMI	Vinh et al. (2009, 2010)

[*] TP_i (true positive), FP_i (false positive), TN_i (true negative) and FN_i (false negative) are diagnoses of how well the data points in the *i*th predicted cluster compared with the best matching real cluster.

[ˆ] TP (true positive), FP (false positive), TN (true negative) and FN (false negative) refers to diagnoses of pairs of data points that were clustered into the same or different clusters when comparing the real and predicted clusters.

[§] Entropy is calculated as $\sum_{i=1}^{k}P_i\log_2(P_i)$, P_i is the ratio of points falling in cluster i to points not in cluster i.

adjusted Rand statistic (also known as the normalised Rand statistic), proposed by Hubert and Arabie (1985), was perhaps one of the most popular indexes used in practice (Rodriguez et al., 2019; Yeung and Ruzzo, 2001). Similar normalization can also be obtained for other measures with counting pairs data points such as FM and Jaccard coefficient, which were found to have very similar validation performances as the adjusted Rand statistic (Aggarwal and Reddy, 2013). Detailed comparisons of different external criteria can be found elsewhere (Amigó et al., 2009).

5.1.2. *Internal criteria*

Internal criteria evaluate specifically whether the clustering model describes the underlying data accurately (goodness-of-fit indicators). A commonly used method is the Cophenetic Correlation Coefficient (CPCC) for hierarchical clustering, which measures the correlation between the input distance (dissimilarity measures between points) and output distance on the dendrogram (Sokal and Rohlf, 1962). Although widely used, the CPCC should be interpreted with care as it is not a direct measure of goodness-of-fit and is sensitive to outliers, nonlinear associations and lower levels of separations in clusters (Farris, 1969; Holgersson, 1978; Mérigot et al., 2010).

5.1.3. *Relative criteria*

Relative criteria have received considerable attention in the literature as key elements in choosing the best number of clusters and validating clustering results. A variety of measures being developed are based on measurements of clustering compactness (homogeneity within the cluster), separation (between cluster distance), representativeness (representative of the underlying data structure), connectedness (clustered similarity with nearest neighbors), stability (consistency in results with subgroups of data) and various combination of these features. The commonly adopted methods were listed in Table 6.

Criteria based on compactness, such as root-mean-square standard deviation (RMSSD), works well for spherical and well-separated clusters

Table 6
Relative Validation Indexes and Implementation in R.

Indexes	Equation *k :number of clusters* *j: a data point* *n: number of data points* C_i cluster i	Note	R-packages	References
Root-mean-square standard deviation (RMSSTD)	$\text{RMSSTD} = \sqrt{\frac{\sum_{i=1}^{k} SS_i}{\sum_{i=1}^{k} df_i}}$ $SS_i = \sum_{j=1}^{n} (x_j - \bar{x}_i)^2$ $df_i = No.\ in$ cluster $i - 1$	Square root of the pooled individual clusters variance (SS_i). It measures the within cluster homogeneity.	Can be directly calculated	(Halkidi et al., 2002)
R-squared	$\text{R}^2 = \frac{SS_t - SS_w}{SS_t}$ $SS_t = \sum_{j=1}^{n} (x_j - \bar{x})^2$ $SS_w = \sum_{i=1}^{k} \sum_{\in i} (x_j - \bar{x}_i)^2$ $\bar{x}$ is the mean of all data $\bar{x}_i$ is the mean of cluster i	Measures the ratio of sum of squares between clusters ($SS_t - SS_w$) to the total sum of squares (SS_t). It measures the degrees of separation between clusters.	Can be directly calculated	(Halkidi et al., 2002)
Normalized Hubert Γ statistic	$\Gamma = \frac{\sum_{i}^{n-1} \sum_{j=i+1}^{n} (P_{ij} - \mu_P)(Q_{ij} - \mu_Q)}{M\sigma_P\ \sigma_Q}$ $M = \frac{n(n-1)}{2}$ P is the distance matrix of data point, Q is the matrix of cluster distances which individual points belong to. μ_P,$\mu_{Q,}\sigma_P$ and $\sigma_Q$are the respective means and variances of the P and Q matrices.	Measures the correlation between data points and their representing clusters. The higher Normalized Hubert Γ indicate the existence of compact clusters.	NbClust:: NbClust	(Halkidi et al., 2002)
Calinski–Harabasz (CH) Index	$\text{CH} = \frac{(n-k)B}{(k-1)W}$ $B = \sum_{i=1}^{k} n_i d(C_i, C)^2$ $W = \sum_{i=1}^{k} \sum_{j \in C_i} d(j, C_i)^2$ $d(C_i, C)$ is the distance between cluster C_i to the center of all data, $d(j, C_i)$ is the distance between data point j and its cluster center C_i.	Based on the average between- (B) and within-cluster (W) sum of squares. Normally give preferences to convex shape clusters and do not work well with arbitrary shapes.	fcp::cluster. stats	(Caliński and Harabasz, 1974)
Dunn index	$\text{Dunn} = \frac{\underset{1 \le i < j \le k}{min}\ d(C_i,\ C_j)}{\underset{1 \le g \le k}{max} \text{diam}(\text{C}_g)}$ $d(C_i,\ C_j)$ is the dissimilarity function between two clusters C_i and C_j; $\text{diam}(\text{C}_g)$ is the diameter of the cluster C_g. Both $d(C_i,\ C_j)$ and $\text{diam}(\text{C}_g)$ can be measured in a variety of ways.	Ratio between the minimal between-cluster distance, $d(C_i,\ C_j)$, to maximal within-cluster distance $\text{diam}(\text{C}_g)$. It can be time-consuming to estimate and can be sensitive to noise.	fcp::cluster. stats clValid::dunn	(Dunn, 1974)
Davies–Bouldin (DB) index	$\text{BD} = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(C_i, C_j)} \right)$ $d(C_i, C_j)$ is the distance between centroids of cluster C_i and C_j. $\sigma_i = \sqrt{\frac{1}{n_i} \sum_{x \in i} (x - C_i)^2}$ is the standard deviation of the distance of data points in cluster i.	Sum ratio of within-cluster scatter to between-cluster separation. A lower DB index relates to a model with better separation between the clusters. Similar to CH index, it does not work well with arbitrary shapes.	clusterSim:: index.DB	(Davies and Bouldin, 1979)
Silhouette index	$\text{Silhouette} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{x \in C_i} \frac{b(x) - a(x)}{\max[a(x); b(x)]}$ $a(x) = \frac{1}{n_i - 1} \sum_{y \in C_i\ y \neq x} d(x, y)$ is average distance of data point x with all other data points in same cluster $b(x) = \underset{j \neq i}{min} \left[\frac{1}{n_j} \sum_{y \in C_j} d(x, y) \right]$ is the average distance of x with all data points in the closest cluster.	Sum of pairwise difference of between-cluster distances, $b(x)$, and within-cluster distances $a(x)$. Higher values indicate better clustering results. It is computationally intensive to estimate.	cluster:: silhouette	(Rousseeuw, 1987)
SD index [a]	$\text{SD} = \alpha \times Scat(k) + Dis(k)$ $\alpha =\ Dis(k_{max})$ is a weighting factor. $Scat(k) = \frac{1}{k} \sum_{i=1}^{k} \frac{\parallel \sigma(v_i) \parallel}{\parallel \sigma(v) \parallel}$	Linear combination of average scattering of clusters, $Scat(k)$ and total separation between clusters $Dis(k)$. $Scat(k)$ is calculated as the ratio of cluster variance to data set variance. $Dis(k)$ estimates the separation based on the distances between cluster centers.	clv::clv.SD	(Halkidi et al., 2000)

(*continued on next page*)

Table 6 (*continued*)

Indexes	Equation *k :number of clusters* *j: a data point* *n: number of data points* C_i cluster i	Note	R-packages	References
	$\sigma(v_i)$ is the variance vector for each variable in the cluster i and $\sigma(v)$ vector of variances in all dataset. $Dis(k) = \frac{D_{max}}{D_{min}} \sum_{i=1}^{k} (\sum_{j=1}^{k} \parallel C_i - C_j \parallel)^{-1}$ D_{max} and D_{min} the maximum and minimum distance between cluster centers			
SDbw index	$\mathrm{SDbw} = Scat(k) + \ Den(k)$ $Scat(k)$ is defined the same as SD index. $Den(k) \ = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} \frac{den(C_i \ \cup \ C_j)}{\max(den(C_i), \ den(C_j))}$ $den(C_i) \ = \sum_{x \in C_i} f(x, u_i)$ u_i is the center of C_i $f(x, u_i) = \{ \begin{matrix} 0 \ if \ d(x, u_i) > stdv \\ 1 \ otherwise \end{matrix}$ $stdv$ is the average standard clusters	Introduced the intercluster density measure, $Den(k)$, based on SD index. It evaluates the average density between pairwise clusters, $den(C_i \ \cup \ C_j)$, in relation to the density within clusters, $den(C_i)$ and $den(C_j)$.	clv::clv.SDbw	(Halkidi and Vazirgiannis, 2001)
Clustering Validation index based on Nearest Neighbours (CVNN)	$CVNN(C, \eta) = \frac{Sep(C, \eta)}{\max_{C \in \Gamma}(sep(C, \eta))} + \frac{com(C)}{\max_{C \in \Gamma}(com(C))}$ $sep(C, \eta) = \max_{1 \leq i \leq k} \left(\frac{1}{n_i} \sum_{j \in C_i} \frac{q_\eta(i)}{\eta} \right)$ $com(C)$ is average of all within-cluster dissimilarities. $q_\eta(i)$ is the number of observations among η nearest neighbours that are in other clusters Γ is all possible clustering results compared.	CVNN is based on the intercluster separation, $sep(C, \eta)$, and intracluster compactness, $Com(C, \eta)$. $sep(C, \eta)$ measures the level of overlap (with points' nearest neighbours) in the highest overlapping cluster. The two terms were normalised before adding them up. Smaller values indicate better clustering.	fcp::cvnn	(Liu et al., 2013)
Gap index	$G(k) \ = E_n(\log(W_k)) - \log(W_k)$ $W_k \ = \ \sum_{i=1}^{k} \frac{1}{2n_i} Di$ $D_i \ = \sum_{j, \ j' \in C_i} d_{j, j'}$	W_k measures the expected pooled within-cluster sum of squares around the cluster means. The ideal is to compare W_k obtained from the data with its expectation under a null reference distribution of the data (e.g., uniform distribution). When there are smaller subclusters within large well-separated clusters, it can show non-monotonic behavior, so it is important to evaluate the overall gap curve rather than simply take the optimal value.	cluster::clusGap	(Tibshirani et al., 2001)

[a] $\parallel X \parallel = \sqrt{X^T X}$, X is a column vector.

and will give preferences to algorithms that minimize the within-cluster variation (such as K-means). However, they may fail in detecting more complicated data structures such as arbitrarily shaped clusters. Cluster separation measures how distinct or well-separated the clusters are. Pairwise cluster distances, such as centroid linkage shown in Fig. 4 and R-squared can be used in this framework.

Hubert Γ statistic as well as modified and normalized Hubert Γ statistic (Halkidi et al., 2002) adopt a similar idea to CPCC. They aim at comparing agreement/disagreement between the clustering results with the underlying data structure. As RMSSD, R-squared and Γ statistic are all evaluating one aspect of the clustering criteria, they will all monotonically increase when the number of clusters increases (Xiong and Li, 2013). A common method is to apply the "elbow" method to identify a drastic changing point in the fitting index. However, the "elbow" method is commonly ambiguous and subjective.

As compactness and separation measures both only provide limited information on their own, they are usually combined to form overall indicators of a balanced with-in cluster homogeneity and between-cluster separation. Well known examples are Calinski-Harabasz (CH) index (Caliński and Harabasz, 1974), Dunn index (Dunn, 1974), Davies–Bouldin (DB) index (Davies and Bouldin, 1979) and Silhouette index (Rousseeuw, 1987), see details in Table 6. Most of these indexes give preference to spherical or convex shapes and do not work well with arbitrary shapes. CH Index was also found to be more sensitive to noise in the data (Xiong and Li, 2013). A few recently developed indexes have moved slightly away from evaluating purely compactness and separation. For example, SDbw index proposed by Halkidi and Vazirgiannis (2001) extended the SD index by introducing a density measurement that compares density areas between clusters with density within clusters, considering well-separated clusters should have considerable density decay in the area separating them. Another index that was found to be able to work properly in arbitrarily shaped clusters is the Clustering Validation index based on Nearest Neighbors (CVNN). CVNN evaluates whether data points were grouped into the same clusters as their nearest neighbors, which facilitates the rationale of density-based clustering algorithms.

For model-based clustering, goodness-of-fit indexes, such as Bayesian information criterion (BIC), are commonly used as relative criteria to compare between models (Fraley and Raftery, 1998). Hypothesis testing can also be employed to select the best numbers of clusters, e.g., Vuong-Lo-Mendell-Rubin adjusted likelihood ratio test (Vuong, 1989) and bootstrapped likelihood ratio test (McLachlan, 1987).

These criteria measure different types of clustering quality and should be chosen according to the clustering method(s) used and the features of the data. In practice, users should aim to report results from multiple criteria to obtain a more comprehensive view of clustering performance.

5.2. Resampling and cross-validation

An important validation process (e.g., choosing the best number of clusters) to include for all clustering tasks is resampling and Cross-Validation (CV). This is because many methods are sensitive to small variations in data and random seed and the risk for overfitting is high. Many types of resampling and validation methods have been developed (Fig. 14). All these methods involve creating subsamples of the data to establish models jointly to improve the stability and generalizability of the model as well as to avoid over-fitting the data. The common CV regime involves using both training dataset(s) (e.g., identifying best hyperparameters) and testing dataset(s) (e.g., prediction accuracy). As it is not possible to evaluate prediction accuracy in clustering (unknown cluster membership), CV is commonly used to in hyperparameter selection (e.g., number of clusters) using training datasets. However, testing data can be used to evaluate the generalizability of the established model (e.g., whether the clusters identified in the testing data under the same modelling parameters were consistent with training data).

K-fold CV, Monte Carlo CV and Bootstrap are commonly used for selecting the best modelling parameters (Sylvain and Alain, 2010).

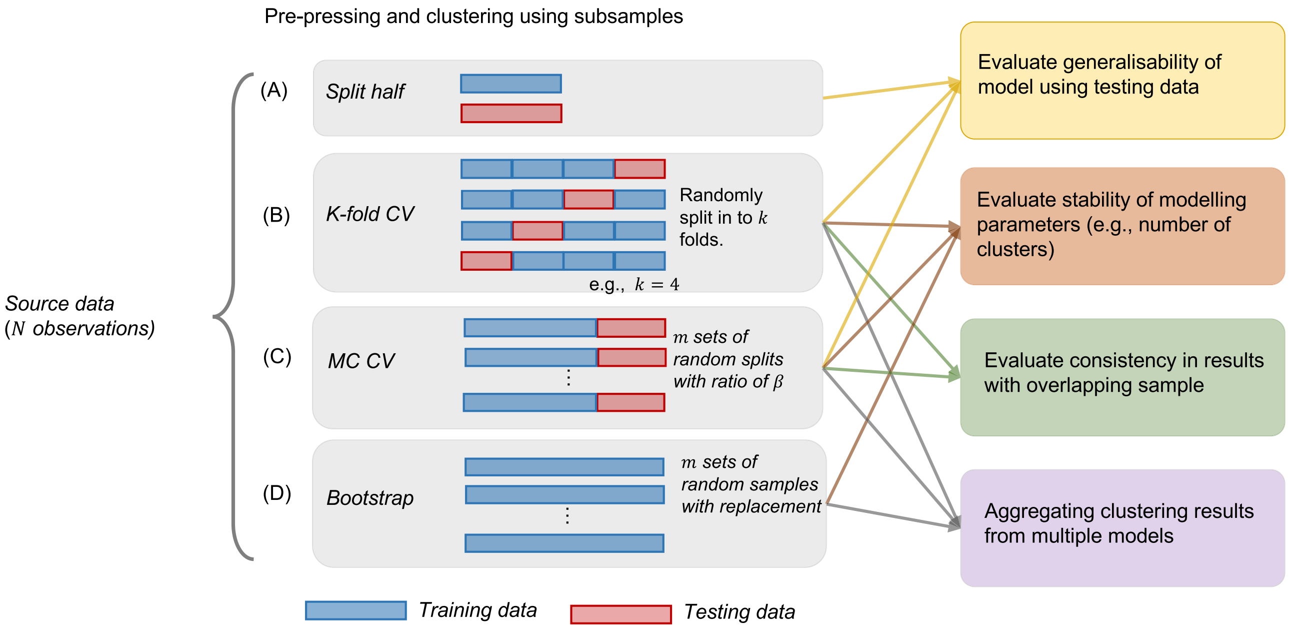

Pre-pressing and clustering using subsamples
Source data
(N observations)
(A)
Split half
(B)
K-fold CV
Randomly split in to k folds.
e.g., $k = 4$
(C)
MC CV
m sets of random splits with ratio of β
(D)
Bootstrap
m sets of random samples with replacement
Evaluate generalisability of model using testing data
Evaluate stability of modelling parameters (e.g., number of clusters)
Evaluate consistency in results with overlapping sample
Aggregating clustering results from multiple models
Training data
Testing data

Fig. 14. Common Methods for Resampling and Cross Validation (CV).
Note: (A) The total sample can be randomly split into half with one being the "training data" and one being the "testing data". The optimal model can be obtained in the training data, and then evaluated in the testing data. (B) In k-fold CV, the total dataset is randomly split into k-fold with equal size. Then the data can be organised into k pairs of training (k-1 folds) and testing datasets (1-fold). The commonly used method is the 10-fold CV. When k is the same as the sample size, the method became leave-one-out CV. (C) Monte Carlo CV, repetitively randomly samples a proportion of data into the training data and leaves the remaining data for testing. When the population is large, a much smaller number of training and testing samples can be selected. (D) Bootstrap creates random samples using sampling with replacement. Therefore, one observation can be sampled into a resampled dataset multiple times.

These methods have different assumptions, advantages, and disadvantages; however, they can yield comparable results when properly specified (Molinaro et al., 2005). Bootstrap can retain the same sample size, establish confidence limits for hard clustering methods (Suzuki and Shimodaira, 2006) and apply in hypothesis testing (McLachlan, 1987). However, bootstrap samples cannot be used to evaluate agreements between sampled datasets (no one-to-one matching due to sampling with replacement) and can introduce high bias and unstable results (Efron and Tibshirani, 1997; Kohavi, 1995).

K-fold CV is perhaps theoretically more attractive as it explores all possible combinations of small groups of data. As the parameter k is commonly recommended to be between 10 and 20 (Kohavi, 1995), the method is also practically more desirable due to the lower computational cost (Molinaro et al., 2005). However, due to its restricted number of resampling datasets, it may not be a preferred method for clustering ensemble.

There is no optimal CV method for all different practical problems and the performance of CV depends on whether sampled data represent the underlying distribution. The choice of CV model can be based on features of the dataset as well as the overall clustering framework (e.g., whether clustering ensemble is used), and evaluation pipelines can be established to ensure CV leads to stable and generalizable results (e.g., testing with more than one CV methods).

6. Workflow and reporting

Different clustering methods and different procedures employed may result in different partitioning of a data set. The optimal solution is largely dependant on the type of data used and how well the heterogenous groups were reflected by the data. Although there are no gold-standard procedures, here we provide a general guide for the clustering workflow, which could assist with improving the quality, efficiency, and transparency of clustering tasks. The recommended workflow consists of six steps (Fig. 15). It is important to apply the data pre-processing procedures, pretesting, clustering model optimization and visual evaluation within the CV loop (e.g., dimensionality reduction in each resampled dataset) to avoid overfitting and reduce bias. After the clusters were identified, there is a need to further validate their meaningfulness (e.g., evaluate how clustering results correlate with external variables) and external validity (e.g., evaluate whether the identified clusters are generalizable in external data).

To achieve greater research transparency, the analysis plan should be pre-registered and results should be reported with sufficient details, including nature of data, theory supporting possible subgroups, detailed analysis procedures, implementation methods (e.g., software packages, code), validity and generalizability of findings. Although Aldenderfer and Blashfield (1984) established the original framework for reporting clustering results, their recommendations were outdated to meet the

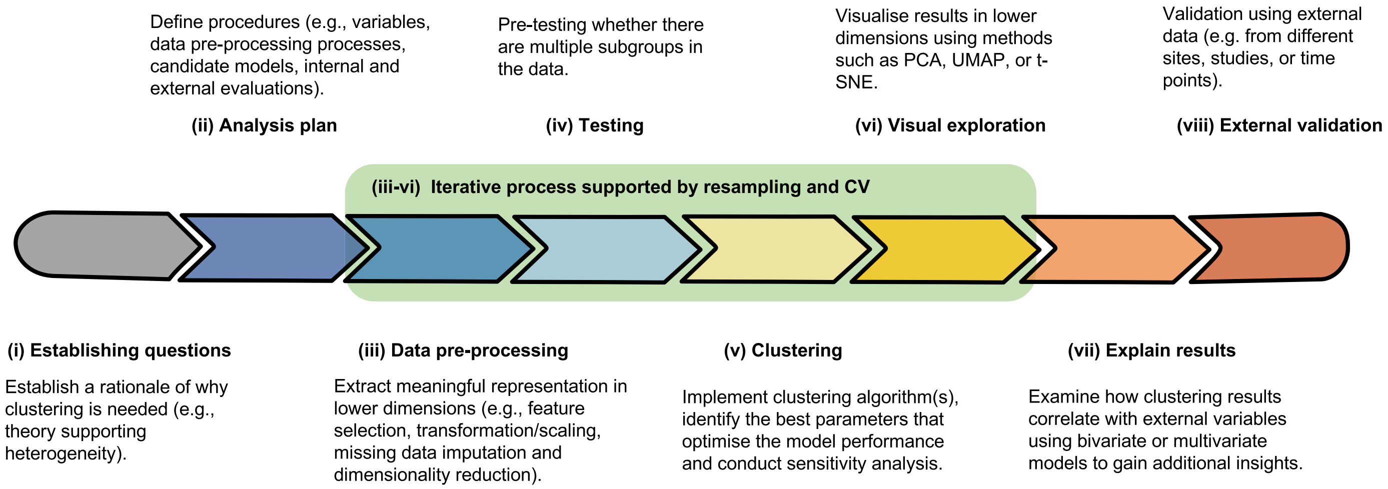

Define procedures (e.g., variables, data pre-processing processes, candidate models, internal and external evaluations).
(ii) Analysis plan
Pre-testing whether there are multiple subgroups in the data.
(iv) Testing
Visualise results in lower dimensions using methods such as PCA, UMAP, or t-SNE.
(vi) Visual exploration
Validation using external data (e.g. from different sites, studies, or time points).
(viii) External validation
(iii-vi) Iterative process supported by resampling and CV
(i) Establishing questions
Establish a rationale of why clustering is needed (e.g., theory supporting heterogeneity).
(iii) Data pre-processing
Extract meaningful representation in lower dimensions (e.g., feature selection, transformation/scaling, missing data imputation and dimensionality reduction).
(v) Clustering
Implement clustering algorithm(s), identify the best parameters that optimise the model performance and conduct sensitivity analysis.
(vii) Explain results
Examine how clustering results correlate with external variables using bivariate or multivariate models to gain additional insights.

Fig. 15. Clustering Workflow.
Note: PCA: Principal Component Analysis; UMAP: Uniform Manifold Approximation; t-SNE: T-distributed Stochastic neighbor Embedding.

Table 7
Check-list for pre-registration and reporting clustering analysis.

	Pre-registration	Final publication
Nature of the data and variables	Context	Context & results
Rationale/theory supporting clustering analysis (e.g., theory supporting possible subgroups which can be identified using selected variables)	Context	Context & discussion
Data pre-processing procedures	Methods	Methods & results
Similarity/distance measure(s) and justification	Methods if used	Methods if used
Pre-clustering testing	Methods	Methods & results
Clustering method(s) and justification	Methods	Methods & results
Selecting modelling parameters	Methods	Methods & results
Missing data and methods to deal with missingness	Methods	Methods & missingness
Resampling, CV, and/or external validation	Methods	Methods & results
Visual representation of clustering results	Methods	Methods & results
Evaluation of cluster meaningfulness	Methods	Methods & results
Computer program(s), package(s), function(s) and associated version	Not necessary	All details
Sensitivity analysis	Methods	Methods & results
Deviation from analysis plan	–	Methods & results
Research data (or synthetically generated data) and analysis code for replication	–	Supplementary files or citable resources

needs of increasing complexity in analysis procedures. Therefore, we proposed an additional checklist for both pre-registration and final publication (Table 7).

7. Summary

Clustering analysis is a longstanding but rapidly evolving machine learning area. Clustering methods are often difficult to choose, justify, robustly conduct, evaluate and validate. There have been many innovative clustering methods developed and/or applied in mental health research, such as FRF (Feczko et al., 2018) and HYDRA (Varol et al., 2017) as mentioned above. Tokuda et al. (2018) developed a multi-view clustering based on a non-parametric Bayesian mixture model for depression subtyping. Chen et al. (2020) applied fuzzy C-means and GMM jointly to identify ambiguous data points which may not belong to any schizophrenia subtypes. Dwyer et al. (2020) used a consensus clustering based on nonnegative matrix factorization to evaluate psychosis subgroups. To deal with the high dimensionality issue in neural imagining data, Chang and colleagues combined deep autoencoder with clustering ensembles (Chang et al., 2021). Most of these advanced modelling approaches have been applied in neuroscience. The current standard practice of using clustering models in mental health research remains relatively simple and lacks a robust framework. Across fields, research is also largely limited to describing empirical findings. We hope our overview and recommendations can assist mental health researchers to use these methods efficiently, transparently and robustly to produce results that lead towards clinical use and theoretical understanding of principles underlying illness.

CRediT authorship contribution statement

Caroline X. Gao: Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Dominic Dwyer:** Conceptualization, Methodology, Writing – review & editing. **Ye Zhu:** Conceptualization, Methodology, Writing – review & editing. **Catherine L. Smith:** Methodology, Writing – review & editing. **Lan Du:** Methodology, Writing – review & editing. **Kate M. Filia:** Methodology, Writing – review & editing. **Johanna Bayer:** Methodology, Writing – review & editing. **Jana M. Menssink:** Methodology, Writing – review & editing. **Teresa Wang:** Methodology, Writing – review & editing. **Christoph Bergmeir:** Methodology, Writing – review & editing. **Stephen Wood:** Conceptualization, Methodology, Writing – review & editing. **Sue M. Cotton:** Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

This study was not funded. The author(s) declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgments

The designs of some of the graphical illustrations were inspired by online posts and blogs from the generous machine learning community. A/Prof. Lianne Schmaal, Dr James Kean, Dr Sarah Herniman and other researchers from the Orygen Clinical Neuroscience Group and Monash Biostatistics Unit have all provided valuable feedback on the contents.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2023.115265.

References

Abi-Dargham, A., Horga, G., 2016. The search for imaging biomarkers in psychiatric disorders. Nat. Med. 22 (11), 1248–1255. https://doi.org/10.1038/nm.4190.

Abramovitch, A., Short, T., Schweiger, A., 2021. The C Factor: cognitive dysfunction as a transdiagnostic dimension in psychopathology. Clin. Psychol. Rev. 86, 102007 https://doi.org/10.1016/j.cpr.2021.102007.

Adolfsson, A., Ackerman, M., Brownstein, N.C., 2019. To cluster, or not to cluster: an analysis of clusterability methods. Pattern Recognit. 88, 13–26. https://doi.org/10.1016/j.patcog.2018.10.026.

Aggarwal, C.C., Hinneburg, A., Keim, D.A., 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In: Van den Bussche, J., Vianu, V. (Eds.), Database Theory ICDT 2001. ICDT 2001. Lecture Notes in Computer Science, vol 1973. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44503-X_27.

Aggarwal, C.C., Reddy, C.K., 2013. Data Clustering: Algorithms and Applications. CRC Press LLC. http://ebookcentral.proquest.com/lib/monash/detail.action?docID=1355921.

Aghabozorgi, S., Seyed Shirkhorshidi, A., Ying Wah, T., 2015. Time-series clustering – a decade review. Inf. Syst. 53, 16–38. https://doi.org/10.1016/j.is.2015.04.007.

Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 1998. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Seattle, Washington, USA. https://doi.org/10.1145/276304.276314.

Aizerman, M.A., 1964. Theoretical foundations of the potential function method in pattern recognition learning. Autom. Remote Control 25, 821–837.

Alamuri, M., Surampudi, B.R., Negi, A., 2014. A survey of distance/similarity measures for categorical data. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 1907–1914. https://doi.org/10.1109/IJCNN.2014.6889941.

Aldenderfer, M.S., Blashfield, R.K., 1984. Cluster Analysis. Sage Publications, CA.

Amigó, E., Gonzalo, J., Artiles, J., Verdejo, M., 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inform Retriev 12:461-486 Inf. Retr. Boston 12, 461–486. https://doi.org/10.1007/s10791-008-9066-8.

Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. ACM Sigmod. Record. 28 (2), 49–60.

Arthur, D., & Vassilvitskii, S. (2006). k-means++: The Advantages of Careful Seeding. http://ilpubs.stanford.edu:8090/778/.

Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. Advances in Latent Variable Mixture Models, 27–51.

Bagga, A., Baldwin, B., 1998. Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1. Montreal, Quebec, Canada. https://doi.org/10.3115/980845.980859.

Bair, E., 2013. Semi-supervised clustering methods. Wiley Interdiscip. Rev. Comput. Stat. 5 (5), 349–361. https://doi.org/10.1002/wics.1270.

Bair, E., Tibshirani, R., 2004. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol. 2 (4), E108. https://doi.org/10.1371/journal.pbio.0020108.

Ball, G.H., .& Hall, D.J. (.1965). ISODATA, a Novel Method of Data Analysis and Pattern Classification.

Bandaragoda, T.R., Ting, K.M., Albrecht, D., Liu, F.T., Zhu, Y., Wells, J.R., 2018. Isolation-based anomaly detection using nearest-neighbor ensembles. Comput. Intell. 34 (4), 968–998. https://doi.org/10.1111/coin.12156.

Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L., Rathouz, P.J., 1997. Latent variable regression for multiple discrete outcomes. J. Am. Stat. Assoc. 92 (440), 1375–1386. https://doi.org/10.1080/01621459.1997.10473658.

Basagaña, X., Barrera-Gómez, J., Benet, M., Antó, J.M., Garcia-Aymerich, J., 2013. A framework for multiple imputation in cluster analysis. Am. J. Epidemiol. 177 (7), 718–725. https://doi.org/10.1093/aje/kws289.

Benaglia, T., Chauveau, D., Hunter, D.R., Young, D.S., 2009. mixtools: an R package for analyzing mixture models. J. Stat. Softw. 32 (6), 1–29. https://doi.org/10.18637/jss.v032.i06.

Berndt, D.J., Clifford, J., 1994. Using dynamic time warping to find patterns in time series. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.5555/3001460.3001507.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Springer.

Bhattacharjee, P., Mitra, P., 2020. A survey of density based clustering algorithms. Front. Comput. Sci. 15 (1), 151308 https://doi.org/10.1007/s11704-019-9059-3.

Booij, M.M., van Noorden, M.S., van Vliet, I.M., Ottenheim, N.R., van der Wee, N.J.A., Van Hemert, A.M., Giltay, E.J., 2021. Dynamic time warp analysis of individual symptom trajectories in depressed patients treated with electroconvulsive therapy. J. Affect Disord. 293, 435–443. https://doi.org/10.1016/j.jad.2021.06.068.

Boongoen, T., Iam-On, N., 2018. Cluster ensembles: a survey of approaches with recent extensions and applications. Comput. Sci. Rev. 28, 1–25. https://doi.org/10.1016/j.cosrev.2018.01.003.

Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, Texas, USA. https://doi.org/10.1145/342009.335388.

Brusco, M., Steinley, D., Watts, A.L., 2022. A comparison of spectral clustering and the walktrap algorithm for community detection in network psychometrics. Psychol. Methods. https://doi.org/10.1037/met0000509. No Pagination Specified-No Pagination Specified.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3 (1), 1–27. https://doi.org/10.1080/03610927408827101.

Campello, R.J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. Advances in Knowledge Discovery and Data Mining, Berlin, Heidelberg. 10.1007/978-3-642-37456-2_14.

Cardot, H., Cénac, P., Monnez, J.M., 2012. A fast and recursive algorithm for clustering large datasets with k-medians. Comput. Stat. Data Anal. 56 (6), 1434–1449. https://doi.org/10.1016/j.csda.2011.11.019.

Carpenter, W.T., Kirkpatrick, B, 1988. The heterogeneity of the long-term course of schizophrenia. Schizophr. Bull. 14 (4), 645–652. http://www.ncbi.nlm.nih.gov/pubmed/3064288.

Caspi, A., Houts, R.M., Ambler, A., Danese, A., Elliott, M.L., Hariri, A., Harrington, H., Hogan, S., Poulton, R., Ramrakha, S., Rasmussen, L.J.H., Reuben, A., Richmond-Rakerd, L., Sugden, K., Wertz, J., Williams, B.S., Moffitt, T.E., 2020. Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the Dunedin birth cohort study. JAMA Netw. Open 3 (4), e203221. https://doi.org/10.1001/jamanetworkopen.2020.3221.

Caspi, A., Houts, R.M., Belsky, D.W., Goldman-Mellor, S.J., Harrington, H., Israel, S., Meier, M.H., Ramrakha, S., Shalev, I., Poulton, R., Moffitt, T.E., 2014. The p factor: one general psychopathology factor in the structure of psychiatric disorders? Clin. Psychol. Sci. 2 (2), 119–137. https://doi.org/10.1177/2167702613497473.

Caspi, A., Moffitt, T.E., 2018. All for one and one for all: mental disorders in one dimension. Am. J. Psychiatry 175 (9), 831–844. https://doi.org/10.1176/appi.ajp.2018.17121383.

Cha, S.H., 2007. Comprehensive survey on distance/similarity measures between probability density functions. Int. J. Math. Models Methods Appl. Sci. 1 (2), 1. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.8446&rep=rep1&type=pdf.

Chand, G.B., Dwyer, D.B., Erus, G., Sotiras, A., Varol, E., Srinivasan, D., Doshi, J., Pomponio, R., Pigoni, A., Dazzan, P., Kahn, R.S., Schnack, H.G., Zanetti, M.V., Meisenzahl, E., Busatto, G.F., Crespo-Facorro, B., Pantelis, C., Wood, S.J., Zhuo, C., Davatzikos, C., 2020. Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning. Brain 143 (3), 1027–1038. https://doi.org/10.1093/brain/awaa025.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: a survey. ACM Comput. Surv. 41 (3), 15. https://doi.org/10.1145/1541880.1541882. Article.

Chang, M., Womer, F.Y., Gong, X., Chen, X., Tang, L., Feng, R., Dong, S., Duan, J., Chen, Y., Zhang, R., Wang, Y., Ren, S., Wang, Y., Kang, J., Yin, Z., Wei, Y., Wei, S., Jiang, X., Xu, K., Wang, F., 2021. Identifying and validating subtypes within major psychiatric disorders based on frontal–posterior functional imbalance via deep learning. Mol. Psychiatry 26 (7), 2991–3002. https://doi.org/10.1038/s41380-020-00892-3.

Chao, G., Sun, S., Bi, J., 2021. A survey on multiview clustering. IEEE Trans. Artif. Intell. 2 (2), 146–168. https://doi.org/10.1109/TAI.2021.3065894.

Chao, G., Wang, S., Yang, S., Li, C., Chu, D., 2022. Incomplete multi-view clustering with multiple imputation and ensemble clustering. Appl. Intell. 52 (13), 14811–14821. https://doi.org/10.1007/s10489-021-02978-z.

Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2014). Multivariate analysis of mixed data: the R Package PCAmixdata. arXiv. 10.48550/arXiv.1411.4911.

Chen, J., Patil, K.R., Weis, S., Sim, K., Nickl-Jockschat, T., Zhou, J., Aleman, A., Sommer, I.E., Liemburg, E.J., Hoffstaedter, F., Habel, U., Derntl, B., Liu, X., Fischer, J.M., Kogler, L., Regenbogen, C., Diwadkar, V.A., Stanley, J.A., Riedl, V., Outcome Survey, I., 2020. Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: an international machine learning study. Biol. Psychiatry 87 (3), 282–293. https://doi.org/10.1016/j.biopsych.2019.08.031.

Chi, J.T., Chi, E.C., Baraniuk, R.G., 2016. k-POD: a method for k-means clustering of missing data. Am. Stat. 70 (1), 91–99. https://doi.org/10.1080/00031305.2015.1086685.

Chiu, D.S., Talhouk, A., 2018. diceR: an R package for class discovery using an ensemble driven approach. BMC Bioinform. 19 (1), 11. https://doi.org/10.1186/s12859-017-1996-y.

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., Horne, R., 2005. The use and reporting of cluster analysis in health psychology: a review. Br. J. Health Psychol. 10 (3), 329–358. https://doi.org/10.1348/135910705X25697.

Cole, V.T., Apud, J.A., Weinberger, D.R., Dickinson, D., 2012. Using latent class growth analysis to form trajectories of premorbid adjustment in schizophrenia. J. Abnorm. Psychol. 121 (2), 388–395. https://doi.org/10.1037/a0026922.

Collins, L.M., Lanza, S.T., 2009. Latent Class and Latent Transition Analysis: With applications in the Social, Behavioral, and Health Sciences, 718. John Wiley & Sons.

Cotton, S.M., Hamilton, M.P., Filia, K., Menssink, J.M., Engel, L., Mihalopoulos, C., Rickwood, D., Hetrick, S.E., Parker, A.G., Herrman, H., Telford, N., Hickie, I., McGorry, P.D., Gao, C.X., 2022. Heterogeneity of quality of life in young people attending primary mental health services. Epidemiol. Psychiatr. Sci. 31, e55. https://doi.org/10.1017/S2045796022000427.

Croon, M., 1990. Latent class analysis with ordered latent classe. Br. J. Math Stat. Psychol. 43 (2), 171–192. https://doi.org/10.1111/j.2044-8317.1990.tb00934.x.

Cunningham, J.P., Ghahramani, Z., 2015. Linear dimensionality reduction: survey, insights, and generalizations. J. Mach. Learn. Res. 16 (1), 2859–2900.

Dalmaijer, E.S., Nord, C.L., Astle, D.E., 2022. Statistical power for cluster analysis. BMC Bioinform. 23 (1), 205. https://doi.org/10.1186/s12859-022-04675-1.

Dara, S., Tumma, P., 2018. Feature extraction by using deep learning: a survey. In: Proceedings of the International Conference on Electronics, Communication and Aerospace Technology (ICECA). https://doi.org/10.1109/ICECA.2018.8474912.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. (2), 224–227.

Day, N.E., 1969. Estimating the components of a mixture of normal distributions. Biometrika 56 (3), 463–474.

De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L., 2000. The Mahalanobis distance. Chemom. Intell. Lab. Syst. 50 (1), 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39 (1), 1–22 (Methodological).

Dhillon, I.S., Guan, Y., Kulis, B., 2004a. Kernel k-means: spectral clustering and normalized cuts. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA. https://doi.org/10.1145/1014052.1014118.

Dhillon, I.S., Guan, Y., Kulis, B., 2004b. A Unified View of Kernel K-Means, Spectral Clustering and Graph Cuts. Citeseer.

Dinga, R., Schmaal, L., Penninx, B.W.J.H., van Tol, M.J., Veltman, D.J., van Velzen, L., Mennes, M., van der Wee, N.J.A., Marquand, A.F., 2019. Evaluating the evidence for biotypes of depression: methodological replication and extension of Drysdale et al. (2017). NeuroImage Clin. 22, 101796 https://doi.org/10.1016/j.nicl.2019.101796.

Dolnicar, S., Grün, B., Leisch, F., Schmidt, K., 2013. Required sample sizes for data-driven market segmentation analyses in tourism. J. Travel Res. 53 (3), 296–306. https://doi.org/10.1177/0047287513496475.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D.J., Etkin, A., Schatzberg, A.F., Sudheimer, K., Keller, J., Mayberg, H.S., Gunning, F.M., Alexopoulos, G.S., Fox, M.D., Pascual-Leone, A., Voss, H.U., Liston, C., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. 23 (1), 28–38. https://doi.org/10.1038/nm.4246.

Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. J. Cybern. 4 (1), 95–104.

Dwyer, D.B., Buciuman, M.O., Ruef, A., Kambeitz, J., Sen Dong, M., Stinson, C., Kambeitz-Ilankovic, L., Degenhardt, F., Sanfelici, R., Antonucci, L.A., Lalousis, P.A., Wenzel, J., Urquijo-Castro, M.F., Popovic, D., Oeztuerk, O.F., Haas, S.S., Weiske, J., Hauke, D., Neufang, S., Koutsouleris, N., 2022. Clinical, brain, and multilevel clustering in early psychosis and affective stages. JAMA Psychiatry 79 (7), 677–689. https://doi.org/10.1001/jamapsychiatry.2022.1163.

Dwyer, D.B., Kalman, J.L., Budde, M., Kambeitz, J., Ruef, A., Antonucci, L.A., Kambeitz-Ilankovic, L., Hasan, A., Kondofersky, I., Anderson-Schmidt, H., Gade, K., Reich-Erkelenz, D., Adorjan, K., Senner, F., Schaupp, S., Andlauer, T.F.M., Comes, A.L., Schulte, E.C., Klöhn-Saghatolislam, F., Koutsouleris, N., 2020. An investigation of psychosis subgroups with prognostic validation and exploration of genetic underpinnings: the PsyCourse study. JAMA Psychiatry 77 (5), 523–533. https://doi.org/10.1001/jamapsychiatry.2019.4910.

Eberle, O., Buttner, J., Krautli, F., Muller, K.R., Valleriani, M., Montavon, G., 2022. Building and interpreting deep similarity models. IEEE Trans. Pattern Anal. Mach. Intell. 44 (3), 1149–1161. https://doi.org/10.1109/TPAMI.2020.3020738.

Edwards, A.W., Cavalli-Sforza, L.L., 1965. A method for cluster analysis. Biometrics 362–375.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap method. J. Am. Stat. Assoc. 92 (438), 548–560. https://doi.org/10.1080/01621459.1997.10474007.

Enders, C.K., Bandalos, D.L., 2001. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Struct. Equ. Model. Multidiscip. J. 8 (3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5.

Eppstein, D., Paterson, M.S., Yao, F.F., 1997. On nearest-neighbor graphs. Discrete Comput. Geom. 17 (3), 263–282. https://doi.org/10.1007/PL00009293.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996a. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD'96. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.5555/3001460.3001507.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996b. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise. KDD.

Ezugwu, A.E., Shukla, A.K., Agbaje, M.B., Oyelade, O.N., José-García, A., Agushaka, J.O., 2020. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. Neural Comput. Appl. 33, 6247–6306. https://doi.org/10.1007/s00521-020-05395-4.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S., Bouras, A., 2014. A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans. Emerg. Top. Comput. 2 (3), 267–279. https://doi.org/10.1109/TETC.2014.2330519.

Farahani, F.V., Karwowski, W., Lighthall, N.R., 2019. Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review [Systematic Review]. Front. Neurosci. 13 (585) https://doi.org/10.3389/fnins.2019.00585.

Farris, J.S., 1969. On the cophenetic correlation coefficient. Syst. Zool. 18 (3), 279–285. https://doi.org/10.2307/2412324.

Feczko, E., Balba, N.M., Miranda-Dominguez, O., Cordova, M., Karalunas, S.L., Irwin, L., Demeter, D.V., Hill, A.P., Langhorst, B.H., Grieser Painter, J., Van Santen, J., Fombonne, E.J., Nigg, J.T., Fair, D.A, 2018. Subtyping cognitive profiles in autism spectrum disorder using a functional random forest algorithm. NeuroImage 172, 674–688. https://doi.org/10.1016/j.neuroimage.2017.12.044.

Feczko, E., Fair, D.A., 2020. Methods and challenges for assessing heterogeneity. Biol. Psychiatry 88 (1), 9–17. https://doi.org/10.1016/j.biopsych.2020.02.015.

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A.M., Nigg, J.T., Fair, D.A., 2019. The heterogeneity problem: approaches to identify psychiatric subtypes. Trends Cogn. Sci. 23 (7), 584–601. https://doi.org/10.1016/j.tics.2019.03.009 (Regul. Ed.).

Filippone, M., Camastra, F., Masulli, F., Rovetta, S., 2008. A survey of kernel and spectral methods for clustering. Pattern Recognit. 41 (1), 176–190. https://doi.org/10.1016/j.patcog.2007.05.018.

Fiori, K.L., Antonucci, T.C., Cortina, K.S., 2006. Social network typologies and mental health among older adults. J. Gerontol. Ser. B 61 (1), P25–P32. https://doi.org/10.1093/geronb/61.1.P25.

Fodor, I.K., 2002. A Survey of Dimension Reduction Techniques. Lawrence Livermore National Lab. https://cs.nju.edu.cn/_upload/tpl/01/0b/267/template267/zhouzh.files/course/dm/reading/reading03/fodor_techrep02.pdf.

Forgy, E.W., 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics 21, 768–769. https://ci.nii.ac.jp/naid/10009668881/en/.

Fowlkes, E.B., Mallows, C.L., 1983. A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. 78 (383), 553–569. https://doi.org/10.2307/2288117.

Fraccaro, P., Beukenhorst, A., Sperrin, M., Harper, S., Palmier-Claus, J., Lewis, S., Van der Veer, S.N., Peek, N., 2019. Digital biomarkers from geolocation data in bipolar disorder and schizophrenia: a systematic review. J. Am. Med. Inform. Assoc. 26 (11), 1412–1420. https://doi.org/10.1093/jamia/ocz043.

Fraley, C., Raftery, A.E., 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. 41 (8), 578–588. https://doi.org/10.1093/comjnl/41.8.578.

Fred, A.L.N., Jain, A.K., 2002. Data clustering using evidence accumulation. In: Proceedings of the International Conference on Pattern Recognition. https://doi.org/10.1109/ICPR.2002.1047450.

Fred, A.L.N., Jain, A.K., 2005. Combining multiple clusterings using evidence accumulation. IEEE Trans. Pattern Anal. Mach. Intell. 27 (6), 835–850. https://doi.org/10.1109/TPAMI.2005.113.

Friedrich, L., 2004. FlexMix: a general framework for finite mixture models and latent class regression in R. J. Stat. Softw. 11 (1), 1–18. https://doi.org/10.18637/jss.v011.i08.

Fu, L., Lin, P., Vasilakos, A.V., Wang, S., 2020. An overview of recent multi-view clustering. Neurocomputing 402, 148–161. https://doi.org/10.1016/j.neucom.2020.02.104.

Gan, G., Ma, C., Wu, J., 2020. Data clustering: theory, Algorithms, and Applications. SIAM.

Gaynor, S., Bair, E., 2017. Identification of relevant subtypes via preweighted sparse clustering. Comput. Stat. Data Anal. 116, 139–154. https://doi.org/10.1016/j.csda.2017.06.003.

Giordani, P., Ferraro, M.B., Martella, F., 2020. An Introduction to Clustering with R. Springer.

Girish, D., Singh, V., Ralescu, A., 2019. Unsupervised Clustering Based Understanding of CNN. CVPR Workshops.

Gönen, M., Margolin, A.A., 2014. Localized data fusion for kernel k-means clustering with application to cancer biology. Adv. Neural Inf. Process Syst. 27, 1305–1313.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. http://ebookcentral.proquest.com/lib/monash/detail.action?docID=6287197.

Goodkind, M., Eickhoff, S.B., Oathes, D.J., Jiang, Y., Chang, A., Jones-Hagata, L.B., Ortega, B.N., Zaiko, Y.V., Roach, E.L., Korgaonkar, M.S., Grieve, S.M., Galatzer-Levy, I., Fox, P.T., Etkin, A., 2015. Identification of a common neurobiological substrate for mental illness. JAMA Psychiatry 72 (4), 305–315. https://doi.org/10.1001/jamapsychiatry.2014.2206.

Goodman, L.A., 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61 (2), 215–231. https://doi.org/10.2307/2334349.

Gordon, A.D. (.1996). Null models in cluster validation. From Data to Knowledge, Berlin, Heidelberg.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27 (4), 857–871. https://doi.org/10.2307/2528823.

Green, M.J., Girshkin, L., Kremerskothen, K., Watkeys, O., Quidé, Y., 2020. A systematic review of studies reporting data-driven cognitive subtypes across the psychosis spectrum. Neuropsychol. Rev. 30 (4), 446–460. https://doi.org/10.1007/s11065-019-09422-7.

Griffiths, T.L., Ghahramani, Z., 2011. The Indian buffet process: an introduction and review. J. Mach. Learn. Res. 12 (4). http://jmlr.org/papers/v12/griffiths11a.html.

Grün, B., Leisch, F., 2008. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J. Stat. Softw. 28 (4), 1–35. https://doi.org/10.18637/jss.v028.i04.

Guha, S., Rastogi, R., Shim, K., 1998. CURE: an efficient clustering algorithm for large databases. ACM Sigmod. Record. 27 (2), 73–84.

Guha, S., Rastogi, R., Shim, K., 2000. ROCK: a robust clustering algorithm for categorical attributes. Inf. Syst. 25 (5), 345–366.

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. J. Intell. Inf. Syst. 17 (2), 107–145. https://doi.org/10.1023/A:1012801612483.

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2002. Clustering validity checking methods: part II. ACM Sigmod. Record. 31 (3), 19–27. https://doi.org/10.1145/601858.601862.

Halkidi, M., Vazirgiannis, M., 2001. Clustering validity assessment: finding the optimal partitioning of a data set. In: Proceedings of the IEEE International Conference on Data Mining. https://doi.org/10.1109/ICDM.2001.989517.

Han, J., Pei, J., Kamber, M., 2011. Data mining: Concepts and Techniques. Elsevier.

Hartigan, J.A., Hartigan, P.M., 1985. The dip test of unimodality. Ann. Stat. 13 (1), 70–84. http://www.jstor.org/stable/2241144.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a K-means clustering algorithm. J. R. Stat. Soc. Ser. C 28 (1), 100–108. https://doi.org/10.2307/2346830 (Applied Statistics).

He, X., Cai, D., Shao, Y., Bao, H., Han, J., 2011. Laplacian regularized gaussian mixture model for data clustering. IEEE Trans. Knowl. Data Eng. 23 (9), 1406–1418. https://doi.org/10.1109/TKDE.2010.259.

Holgersson, M., 1978. The limited value of cophenetic correlation as a clustering criterion. Pattern Recognit. 10 (4), 287–295. https://doi.org/10.1016/0031-3203(78)90038-9.

Huang, A., 2008. Similarity measures for text document clustering. In: Proceedings of the 6th New Zealand Computer Science Research Student Conference (NZCSRSC2008). Christchurch, New Zealand.

Huang, Z., 1997a. Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD).

Huang, Z., 1997b. A fast clustering algorithm to cluster very large categorical data sets in data mining. DMKD 3 (8), 34–39.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2 (1), 193–218. https://doi.org/10.1007/BF01908075.

Hyman, S.E., 2010. The diagnosis of mental disorders: the problem of reification. Annu. Rev. Clin. Psychol. 6, 155–179. https://doi.org/10.1146/annurev.clinpsy.3.022806.091532.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., Wang, P., 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry 167 (7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379.

Insel, T.R., Cuthbert, B.N., 2015. Brain disorders? Precisely. Science 348 (6234), 499–500. https://doi.org/10.1126/science.aab2358.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 31 (8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999a. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323. https://doi.org/10.1145/331499.331504.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999b. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323. https://doi.org/10.1145/331499.331504.

Jajuga, K., Walesiak, M., 2000. Standardisation of Data Set Under Different Measurement Scales. Springer Berlin Heidelberg, pp. 105–112. https://doi.org/10.1007/978-3-642-57280-7_11.

John, C.R., Watson, D., Russ, D., Goldmann, K., Ehrenstein, M., Pitzalis, C., Lewis, M., Barnes, M., 2020. M3C: monte Carlo reference-based consensus clustering. Sci. Rep. 10 (1), 1816. https://doi.org/10.1038/s41598-020-58766-1.

Johnson, S.C., 1967. Hierarchical clustering schemes. Psychometrika 32 (3), 241–254.

Jolliffe, I., 2022. A 50-year personal journey through time with principal component analysis. J. Multivar. Anal. 188, 104820 https://doi.org/10.1016/j.jmva.2021.104820.

Jung, T., Wickrama, K.A., 2008. An introduction to latent class growth analysis and growth mixture modeling. Soc. Pers. Psychol. Compass 2 (1), 302–317.

Kapur, S., Phillips, A.G., Insel, T.R., 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol. Psychiatry 17 (12), 1174–1179. https://doi.org/10.1038/mp.2012.105.

Karim, M.R., Beyan, O., Zappa, A., Costa, I.G., Rebholz-Schuhmann, D., Cochez, M., Decker, S., 2020. Deep learning-based clustering approaches for bioinformatics. Brief. Bioinform. 22 (1), 393–415. https://doi.org/10.1093/bib/bbz170.

Karypis, G., Han, E.H., Kumar, V., 1999. Chameleon: hierarchical clustering using dynamic modeling. Computer 32 (8), 68–75 (Long Beach Calif).

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: an Introduction to Cluster Analysis, 344. John Wiley & Sons.

Koestler, D.C., Marsit, C.J., Christensen, B.C., Karagas, M.R., Bueno, R., Sugarbaker, D.J., Kelsey, K.T., Houseman, E.A., 2010. Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. Bioinformatics 26 (20), 2578–2585. https://doi.org/10.1093/bioinformatics/btq470.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the International Joint Conference on Artificial Intelligence.

Kotov, R., Foti, D., Li, K., Bromet, E.J., Hajcak, G., Ruggero, C.J., 2016. Validating dimensions of psychosis symptomatology: neural correlates and 20-year outcomes. J. Abnorm. Psychol. 125 (8), 1103–1119. https://doi.org/10.1037/abn0000188.

Kotov, R., Krueger, R.F., Watson, D., 2018. A paradigm shift in psychiatric classification: the hierarchical taxonomy of psychopathology (HiTOP) [10.1002/wps.20478]. World Psychiatry 17 (1), 24–25. https://doi.org/10.1002/wps.20478.

Kotov, R., Krueger, R.F., Watson, D., Achenbach, T.M., Althoff, R.R., Bagby, R.M., Brown, T.A., Carpenter, W.T., Caspi, A., Clark, L.A., Eaton, N.R., Forbes, M.K., Forbush, K.T., Goldberg, D., Hasin, D., Hyman, S.E., Ivanova, M.Y., Lynam, D.R., Markon, K., Zimmerman, M., 2017. The hierarchical taxonomy of psychopathology (HiTOP): a dimensional alternative to traditional nosologies. J. Abnorm. Psychol. 126 (4), 454–477. https://doi.org/10.1037/abn0000258.

Kotov, R., Leong, S.H., Mojtabai, R., Erlanger, A.C.E., Fochtmann, L.J., Constantino, E., Carlson, G.A., Bromet, E.J., 2013. Boundaries of Schizoaffective Disorder: revisiting Kraepelin. JAMA Psychiatry 70 (12), 1276–1286. https://doi.org/10.1001/jamapsychiatry.2013.2350.

Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Prog. Artif. Intell. 5 (4), 221–232. https://doi.org/10.1007/s13748-016-0094-0.

Lam, M., Hill, W.D., Trampush, J.W., Yu, J., Knowles, E., Davies, G., Stahl, E., Huckins, L., Liewald, D.C., Djurovic, S., Melle, I., Sundet, K., Christoforou, A., Reinvang, I., DeRosse, P., Lundervold, A.J., Steen, V.M., Espeseth, T., Räikkönen, K., Lencz, T., 2019. Pleiotropic meta-analysis of cognition, education, and schizophrenia differentiates roles of early neurodevelopmental and adult synaptic pathways. Am. J. Hum. Genet. 105 (2), 334–350. https://doi.org/10.1016/j.ajhg.2019.06.012.

Lampinen, J., Oja, E., 1992. Clustering properties of hierarchical self-organizing maps. J. Math. Imaging Vis. 2 (2), 261–272. https://doi.org/10.1007/BF00118594.

Lawson, R.G., Jurs, P.C., 1990. New index for clustering tendency and its application to chemical problems. J. Chem. Inf. Comput. Sci. 30 (1), 36–41. https://doi.org/10.1021/ci00065a010.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Legendre, P., Legendre, L., 2012. Numerical Ecology. Elsevier.

Leisch, F., 2006. A toolbox for K-centroids cluster analysis. Comput. Stat. Data Anal. 51 (2), 526–544. https://doi.org/10.1016/j.csda.2005.10.006.

Li, Y., Schofield, E., Gönen, M., 2019. A tutorial on Dirichlet Process mixture modeling. J. Math. Psychol. 91, 128–144. https://doi.org/10.1016/j.jmp.2019.04.004.

Li, C., Zhang, Y., 2020. Density peak clustering based on relative density optimization. Math. Probl. Eng. 2020, 2816102. https://doi.org/10.1155/2020/2816102.

Liu, F.T., Ting, K.M., Zhou, Z., 2008. Isolation forest. In: Proceedings of the 8th IEEE International Conference on Data Mining. https://doi.org/10.1109/ICDM.2008.17.

Liu, F.T., .Ting, K.M., .& Zhou, Z.H. (2010, 2010//). On detecting clustered anomalies using SCiForest. Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28 (2), 129–137. https://doi.org/10.1109/TIT.1982.1056489.

Low, D.M., Bentley, K.H., Ghosh, S.S., 2020. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig. Otolaryngol. 5 (1), 96–116. https://doi.org/10.1002/lio2.354.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.

Marin, D., Tang, M., Ayed, I.B., Boykov, Y., 2019. Kernel clustering: density biases and solutions. IEEE Trans. Pattern Anal. Mach. Intell. 41 (1), 136–147. https://doi.org/10.1109/TPAMI.2017.2780166.

Marin, D., Tang, M., Ben Ayed, I., Boykov, Y., 2017. Kernel clustering: density biases and solutions. In: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2017.2780166.

Marquand, A.F., Wolfers, T., Mennes, M., Buitelaar, J., Beckmann, C.F., 2016. Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 1 (5), 433–447. https://doi.org/10.1016/j.bpsc.2016.04.002.

Mathisen, B.M., Aamodt, A., Bach, K., Langseth, H., 2020. Learning similarity measures from data. Prog. Artif. Intell. 9 (2), 129–143. https://doi.org/10.1007/s13748-019-00201-2.

McCutcheon, A.L., 1987. Latent Class Analysis. Sage.

McKusick, V.A., 1969. On lumpers and splitters, or the nosology of genetic disease. Perspect. Biol. Med. 12 (2), 298–312. https://doi.org/10.1353/pbm.1969.0039.

McLachlan, G.J., 1987. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. J. R. Stat. Soc. Ser. C 36 (3), 318–324. https://doi.org/10.2307/2347790 (Applied Statistics).

Meng, T., Jing, X., Yan, Z., Pedrycz, W., 2020. A survey on machine learning for data fusion. Inf. Fusion 57, 115–129. https://doi.org/10.1016/j.inffus.2019.12.001.

Mérigot, B., Durbec, J.P., Gaertner, J.C., 2010. On goodness-of-fit measure for dendrogram-based analyses. Ecology 91 (6), 1850–1859. https://doi.org/10.1890/09-1387.1.

Milligan, G.W., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45 (3), 325–342. https://doi.org/10.1007/BF02293907.

Milligan, G.W., Cooper, M.C., 1987. Methodology review: clustering methods. Appl. Psychol. Meas. 11 (4), 329–354. https://doi.org/10.1177/014662168701100401.

Milligan, G.W., Cooper, M.C., 1988. A study of standardization of variables in cluster analysis. J. Classif. 5 (2), 181–204. https://doi.org/10.1007/BF01897163.

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., Long, J., 2018. A survey of clustering with deep learning: from the perspective of network architecture. IEEE Access 6, 39501–39514. https://doi.org/10.1109/ACCESS.2018.2855437.

Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21 (15), 3301–3307. https://doi.org/10.1093/bioinformatics/bti499.

Monti, S., Tamayo, P., Mesirov, J., Golub, T., 2003. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn. 52 (1), 91–118. https://doi.org/10.1023/A:1023949509487.

Müllner, D., 2013. Fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. J. Stat. Softw. 53 (9), 1–18. https://doi.org/10.18637/jss.v053.i09.

Murtagh, F., Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. Wires Data Min. Knowl. Discov. 2 (1), 86–97. https://doi.org/10.1002/widm.53.

Muthén, B., Asparouhov, T., 2020. Latent transition analysis with random intercepts (RI-LTA). Psychol. Methods. https://doi.org/10.1037/met0000370. No Pagination Specified-No Pagination Specified.

Ng, R.T., Han, J., 2002. CLARANS: a method for clustering objects for spatial data mining. IEEE Trans. Knowl. Data Eng. 14 (5), 1003–1016.

Manduchi, L., Chin-Cheong, K., Michel, H., Wellmann, S., & Vogt, J. (2021). Deep conditional Gaussian mixture model for constrained clustering. arXiv. doi: 10.48550/arXiv.2106.06385.

Norouzi, M., Fleet, D.J., .& Salakhutdinov, R.R. (.2012). Hamming distance metric learning. Advances in Neural Information Processing Systems, http://www.cs.utoronto.ca/~norouzi/research/papers/hdml.pdf.

Nunes, A., Trappenberg, T., Alda, M., 2020. The definition and measurement of heterogeneity. Transl. Psychiatry 10 (1), 299. https://doi.org/10.1038/s41398-020-00986-0.

Nutakki, G.C., Abdollahi, B., Sun, W., Nasraoui, O., 2019. An Introduction to Deep Clustering. In: Nasraoui, O., Ben N'Cir, CE. (Eds.), Clustering Methods for Big Data Analytics. Unsupervised and Semi-Supervised Learning. Springer, Cham.

Oberski, D., 2016. Mixture models: latent profile and latent class analysis. Modern Statistical Methods for HCI. Springer, pp. 275–287.

Pagès, J., 2014. Multiple Factor Analysis by Example Using R. Chapman & Hall/CRC. https://doi.org/10.1201/b17700.

Pantelis, C., Velakoulis, D., McGorry, P.D., Wood, S.J., Suckling, J., Phillips, L.J., Yung, A.R., Bullmore, E.T., Brewer, W., Soulsby, B., Desmond, P., McGuire, P.K., 2003. Neuroanatomical abnormalities before and after onset of psychosis: a cross-sectional and longitudinal MRI comparison. Lancet 361 (9354), 281–288. https://doi.org/10.1016/S0140-6736(03)12323-9.

Pattanodom, M., Iam-On, N., Boongoen, T., 2016. Clustering data with the presence of missing values by ensemble approach. In: Proceedings of the Asian Conference on Defence Technology (ACDT). https://doi.org/10.1109/ACDT.2016.7437660.

Pearson, K., 1894. Contributions to the mathematical theory of evolution. Philos. Trans. R. Soc. Lond. A 185, 71–110.

Pinto, J.V., Moulin, T.C., Amaral, O.B., 2017. On the transdiagnostic nature of peripheral biomarkers in major psychiatric disorders: a systematic review. Neurosci. Biobehav. Rev. 83, 97–108. https://doi.org/10.1016/j.neubiorev.2017.10.001.

Preoţiuc-Pietro, D., Sap, M., Schwartz, H.A., Ungar, L., 2015. Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. https://aclanthology.org/W15-1205.pdf.

Qin, X., Ting, K.M., Zhu, Y., Lee, V.C.S., 2019a. Nearest-neighbour-induced isolation similarity and its impact on density-based clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, 33, pp. 4755–4762. https://doi.org/10.1609/aaai.v33i01.33014755.

Qin, Y., Ding, S., Wang, L., Wang, Y., 2019b. Research progress on semi-supervised clustering. Cognit. Comput. 11 (5), 599–612. https://doi.org/10.1007/s12559-019-09664-w.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66 (336), 846–850. https://doi.org/10.2307/2284239.

Reddy, C.K., Vinzamuri, B., 2018. A survey of partitional and hierarchical clustering algorithms. Data Clustering. Chapman and Hall/CRC, pp. 87–110.

Reef, J., Diamantopoulou, S., van Meurs, I., Verhulst, F.C., van der Ende, J., 2011. Developmental trajectories of child to adolescent externalizing behavior and adult DSM-IV disorder: results of a 24-year longitudinal study. Soc. Psychiatry Psychiatr. Epidemiol. 46 (12), 1233–1241. https://doi.org/10.1007/s00127-010-0297-9.

Rodriguez, A., Laio, A., 2014. Clustering by fast search and find of density peaks. Science 344 (6191), 1492–1496. https://doi.org/10.1126/science.1242072.

Robitzsch, A., 2020. Regularized latent class analysis for polytomous item eesponses: an application to SPM-LS data. J. Intell. 8 (3) https://doi.org/10.3390/jintelligence8030030.

Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.d.F., Rodrigues, F.A., 2019. Clustering algorithms: a comparative approach. PLoS One 14 (1), e0210236. https://doi.org/10.1371/journal.pone.0210236.

Romer, A.L., Elliott, M.L., Knodt, A.R., Sison, M.L., Ireland, D., Houts, R., Ramrakha, S., Poulton, R., Keenan, R., Melzer, T.R., Moffitt, T.E., Caspi, A., Hariri, A.R., 2021. Pervasively thinner neocortex as a transdiagnostic feature of general psychopathology. Am. J. Psychiatry 178 (2), 174–182. https://doi.org/10.1176/appi.ajp.2020.19090934.

Ros, F., Guillaume, S., 2019. A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. Expert Syst. Appl. 128, 96–108. https://doi.org/10.1016/j.eswa.2019.03.031.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Rovetta, S., Masulli, F., 2019. Soft Clustering: Why and How-To. Fuzzy Logic and Applications, Cham.

Rui, X., Wunsch, D., 2005. Survey of clustering algorithms. IEEE Trans. Neural Netw. 16 (3), 645–678. https://doi.org/10.1109/TNN.2005.845141.

Russell, S., Norvig, P., 2021. Artificial Intelligence: a Modern Approach, Global Edition. Pearson Education, Limited. http://ebookcentral.proquest.com/lib/monash/detail.action?docID=6563563.

Sander, J., Ester, M., Kriegel, H.P., Xu, X., 1998. Density-based clustering in spatial databases: the algorithm gdbscan and its applications. Data Min Knowl Discov 2 (2), 169–194.

Sato-Ilic, M., Jain, L.C., 2006. Evaluation of fuzzy clustering. M. Sato-Ilic & L. C. Jain Innovations in Fuzzy Clustering: Theory and Applications. Springer Berlin Heidelberg, pp. 105–123. https://doi.org/10.1007/3-540-34357-1_5.

Schork, A.J., Won, H., Appadurai, V., Nudel, R., Gandal, M., Delaneau, O., Revsbech Christiansen, M., Hougaard, D.M., Bækved-Hansen, M., Bybjerg-Grauholm, J., Giørtz Pedersen, M., Agerbo, E., Bøcker Pedersen, C., Neale, B.M., Daly, M.J., Wray, N.R., Nordentoft, M., Mors, O., Børglum, A.D., Werge, T., 2019. A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. Nat. Neurosci. 22 (3), 353–361. https://doi.org/10.1038/s41593-018-0320-0.

Schubert, E., Rousseeuw, P.J., 2019. Faster K-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. Similarity Search and Applications, Cham. https://doi.org/10.1007/978-3-030-32047-8_16.

Schubert, E., Sander, J., Ester, M., Kriegel, H., Xu, X., 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans. Database Syst. 42 (3), 1–21. https://doi.org/10.1145/3068335.

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J 8 (1), 289–317.

Şenbabaoğlu, Y., Michailidis, G., Li, J.Z., 2014. Critical limitations of consensus clustering in class discovery. Sci. Rep. 4 (1), 6207. https://doi.org/10.1038/srep06207.

Sha, Z., Wager, T.D., Mechelli, A., He, Y., 2019. Common dysfunction of large-scale neurocognitive networks across psychiatric disorders. Biol. Psychiatry 85 (5), 379–388. https://doi.org/10.1016/j.biopsych.2018.11.011.

Shen, B., Liu, B.D., Wang, Q., Ji, R., 2014. Robust nonnegative matrix factorization via L<inf>1</inf>norm regularization by multiplicative updating rules. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). https://doi.org/10.1109/ICIP.2014.7026069.

Silverman, B.W., 1981. Using kernel density estimates to investigate multimodality. J. R. Stat. Soc. Ser. B 43 (1), 97–99 (Methodological). http://www.jstor.org/stable/2985156.

Sim, K., Gopalkrishnan, V., Zimek, A., Cong, G., 2013. A survey on enhanced subspace clustering. Data Min. Knowl. Discov. 26 https://doi.org/10.1007/s10618-012-0258-x.

Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andres Terre, H., Shams, Z., Jamnik, M., Liò, P., 2019. Variational autoencoders for cancer data integration: design principles and computational practice. Front. Genet. 10, 1205 https://doi.org/10.3389/fgene.2019.01205.

Sneath, P.H.A., Sokal, R.R., 1973. Numerical Taxonomy. The Principles and Practice of Numerical Classification. W.H. Freeman Co.

Sokal, R.R., 1974. Classification: purposes, principles, progress, prospects. Science 185 (4157), 1115–1123. https://doi.org/10.1126/science.185.4157.1115.

Sokal, R.R., Rohlf, F.J., 1962. The comparison of dendrograms by objective methods. Taxon 11 (2), 33–40.

Sokal, R.R., Sneath, P.H.A., 1963. Principles of Numerical Taxonomy. W.H. Freeman Co.

Sporns, O., 2018. Graph theory methods: applications in brain networks. Dialogues Clin Neurosci 20 (2), 111–121. https://doi.org/10.31887/DCNS.2018.20.2/osporns.

Stan Development Team. (2019). 9.2 Soft K-means. In Stan User's Guide Version 2.27. https://mc-stan.org/docs/2_27/stan-users-guide/soft-k-means.html.

Steinbach, M., Ertöz, L., Kumar, V., Wille, L.T., 2004. The challenges of clustering high dimensional data. New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition. Springer Berlin Heidelberg, pp. 273–309. https://doi.org/10.1007/978-3-662-08968-2_16.

Steinley, D., 2003. Local optima in K-means clustering: what you don't know may hurt you. Psychol. Methods 8 (3), 294–304. https://doi.org/10.1037/1082-989x.8.3.294.

Strehl, A., Ghosh, J., 2002. Cluster ensembles a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617. https://doi.org/10.1162/153244303321897735. Dec.

Suzuki, R., Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22 (12), 1540–1542. https://doi.org/10.1093/bioinformatics/btl117.

Sylvain, A., Alain, C., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4 (none), 40–79. https://doi.org/10.1214/09-SS054.

Tokuda, T., Yoshimoto, J., Shimizu, Y., Okada, G., Takamura, M., Okamoto, Y., Yamawaki, S., Doya, K., 2018. Identification of depression subtypes and relevant brain regions using a data-driven approach. Sci. Rep. 8 (1), 14082. https://doi.org/10.1038/s41598-018-32521-z.

Topchy, A., Jain, A.K., Punch, W., 2004. A mixture model for clustering ensembles. In: Proceedings of the SIAM international conference on data mining. https://doi.org/10.1137/1.9781611972740.35.

Tryon, R., 1939. Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. A. A. Edwards Brothers.

Tueller, S., Lubke, G., 2010. Evaluation of structural equation mixture models: parameter estimates and correct class assignment. Struct. Equ. Model. Multidiscip. J. 17 (2), 165–192. https://doi.org/10.1080/10705511003659318.

Ulbricht, C.M., Chrysanthopoulou, S.A., Levin, L., Lapane, K.L., 2018. The use of latent class analysis for identifying subtypes of depression: a systematic review. Psychiatry Res. 266, 228–246. https://doi.org/10.1016/j.psychres.2018.03.003.

van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B.W.J.H., Waldorp, L.J., Schoevers, R.A., 2015. Association of symptom network structure with the course of depression. JAMA Psychiatry 72 (12), 1219–1226. https://doi.org/10.1001/jamapsychiatry.2015.2079.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., Yau, C., 2021. Bayesian statistics and modelling. Nat. Rev. Methods Primers 1 (1), 1. https://doi.org/10.1038/s43586-020-00001-2.

van der Kloot, W.A., Spaans, A.M., Heiser, W.J., 2005. Instability of hierarchical cluster analysis due to input order of the data: the PermuCLUSTER solution. Psychol. Methods 10 (4), 468–476. https://doi.org/10.1037/1082-989x.10.4.468.

Van Der Maaten, L., Postma, E., Van den Herik, J, 2009. Dimensionality reduction: a comparative. J. Mach. Learn. Res. 10 (66–71), 13. https://members.loria.fr/moberger/Enseignement/AVR/Exposes/TR_Dimensiereductie.pdf.

Varol, E., Sotiras, A., Davatzikos, C., 2017. HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. NeuroImage 145 (Pt B), 346–364. https://doi.org/10.1016/j.neuroimage.2016.02.041.

Vega-Pons, S., Ruiz-Shulcloper, J., 2011. A survey of clustering ensemble algorithms. Int. J. Pattern Recognit. Artif. Intell. 25 (03), 337–372. https://doi.org/10.1142/S0218001411008683.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. Helsinki, Finland. https://doi.org/10.1145/1390156.1390294.

Vinh, N.X., Epps, J., Bailey, J., 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary?. In: Proceedings of the 26th Annual International Conference on Machine Learning.

Vinh, N.X., Epps, J., Bailey, J., 2010. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. 11, 2837–2854. https://www.jmlr.org/papers/volume11/vinh10a/vinh10a.pdf.

Visser, I., Speekenbrink, M., 2010. depmixS4: an R package for hidden Markov models. J. Stat. Softw. 36 (7), 1–21. https://doi.org/10.18637/jss.v036.i07.

Vlachos, M., Kollios, G., Gunopulos, D., 2002. Discovering similar multidimensional trajectories. In: Proceedings of the 18th International Conference on Data Engineering. https://doi.org/10.1109/ICDE.2002.994784.

von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17 (4), 395–416. https://doi.org/10.1007/s11222-007-9033-z.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57 (2), 307–333. https://doi.org/10.2307/1912557.

Wallace, C.S., Dowe, D.L., 2000. MML clustering of multi-state, poisson, von mises circular and gaussian distributions. Stat. Comput. 10 (1), 73–83. https://doi.org/10.1023/A:1008992619036.

Wan, D., Razavi-Far, R., Saif, M., 2020. Cooperative clustering missing data imputation. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC). https://doi.org/10.1109/SMC42975.2020.9283484.

Wang, X., Smith, K., Hyndman, R., 2006. Characteristic-based clustering for time series data. Data Min. Knowl. Discov. 13 (3), 335–364. https://doi.org/10.1007/s10618-005-0039-x.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58 (301), 236–244. https://doi.org/10.1080/01621459.1963.10500845.

Xiong, H., Li, Z., 2013. Clustering Validation Measures. Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, pp. 571–606. https://doi.org/10.1201/9781315373515-23. Data mining and Knowledge Discovery series.

Xu, D., Tian, Y., 2015. A comprehensive survey of clustering algorithms. Ann. Data Sci. 2 (2), 165–193. https://doi.org/10.1007/s40745-015-0040-1.

Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.J., 2007. SCAN: a structural clustering algorithm for networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA. https://doi.org/10.1145/1281192.1281280.

Yang, Y., Wang, H., 2018. Multi-view clustering: a survey. Big Data Min. Anal. 1 (2), 83–107. https://doi.org/10.26599/BDMA.2018.9020003.

Yeung, K.Y., Ruzzo, W.L., 2001. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. Bioinformatics 17 (9), 763–774. http://staff.washington.edu/kayee/pca/supp.ps.

Yin, J., Wang, J., 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, New York, USA. https://doi.org/10.1145/2623330.2623715.

Zaki, M.J., Meira, W., Meira, W., 2014. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.

Zhang, J., Li, C.G., You, C., Qi, X., Zhang, H., Guo, J., Lin, Z., 2019. Self-supervised convolutional subspace clustering network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_Self-Supervised_Convolutional_Subspace_Clustering_Network_CVPR_2019_paper.html.

Zhang, T., Ramakrishnan, R., Livny, M., 1996. BIRCH: an efficient data clustering method for very large databases. ACM Sigmod. Record. 25 (2), 103–114.

Zheutlin, A.B., Dennis, J., Karlsson Linnér, R., Moscati, A., Restrepo, N., Straub, P., Ruderfer, D., Castro, V.M., Chen, C.Y., Ge, T., Huckins, L.M., Charney, A., Kirchner, H.L., Stahl, E.A., Chabris, C.F., Davis, L.K., Smoller, J.W., 2019. Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 106,160 patients across four health care systems. Am. J. Psychiatry 176 (10), 846–855. https://doi.org/10.1176/appi.ajp.2019.18091085.

Zhou, M., Thayer, W.M., Bridges, J.F.P., 2018. Using latent class analysis to model preference heterogeneity in health: a systematic review. Pharmacoeconomics 36 (2), 175–187. https://doi.org/10.1007/s40273-017-0575-4.

Zhu, Y., Ting, K.M., Carman, M.J., Angelova, M., 2021. CDF transform-and-shift: an effective way to deal with datasets of inhomogeneous cluster densities. Pattern Recognit. 117, 107977 https://doi.org/10.1016/j.patcog.2021.107977.

Zouridakis, G., Boutros, N.N., Jansen, B.H., 1997. A fuzzy clustering approach to study the auditory P50 component in schizophrenia. Psychiatry Res. 69 (2–3), 169–181. https://www.sciencedirect.com/science/article/pii/S0165178196029794?via%3Dihub.